Processes and Rates of Bacterial Evolution

Dissertation Advisor: Christopher J. Marx Nigel Delaney

Processes and Rates of Bacterial Evolution

Abstract

A long-standing question in evolutionary biology is whether adaptation will typically proceed through a few mutations with large selective effects or many mutations with small effects. Many studies have implicated few loci of major effect, but it has been predicted that small-effect mutations should exist and contribute to adaptation. However, such mutations have not been found in many studies, either because they do not exist or because the experimental design limited their detection. To determine the effects and types of mutations contributing to adaptation, I studied laboratory and wild populations of bacteria. I characterized the distribution of the effect sizes in laboratory populations of an aerobic bacterium, *Methylobacterium extorquens,* and studied the types of genetic changes associated with adaptation to a novel host in wild populations of *Mycoplasma gallisepticum.*

Chapters 1, 2 and 3 describe tools and strains that I developed to perform extensively replicated evolution experiments in *Methylobacterium extorquens*. I created a robotic system to assay bacterial cultures under high-throughput conditions, designed a medium that allowed for stable growth of *M. extorquens* and genetically manipulated *Methylobacterium* to remove genes preventing reproducible assays in batch culture.

In chapter 4, I applied these tools to infer the distribution of selective effects for beneficial mutations and the rate at which they occur in *M. extorquens*. I evolved 192 populations at two different sizes: a large population size treatment to screen for rare beneficial mutations with large selective effects and a small population size treatment to screen for mutations with smaller

effects. In contrast to expectations, I found a high beneficial mutation rate and a distribution skewed towards large effects.

In chapter 5, I used genome re-sequencing to examine the genetic changes that occurred in a population of pathogenic *Mycoplasma gallisepticum* over a 13-year period following a host shift to house finches (*Carpodacus mexicanus*). In addition to many genomic changes, I observed a surprising loss of the immunity (CRISPR) genes. I recorded the highest rate of nucleotide variation introduced per year measured in a bacterium, demonstrating that one can study the kinetics of genetic change occurring in this population on tractable time-scales.

Table of Contents

Chapter 1

Clarity: An Open-Source Manager for Laboratory Automation

A description of the software created to perform the experiments in this dissertation.

Due to formatting requirements, this work is reprinted in the back matter.

Reprinted from the Journal of Laboratory Automation.

Delaney, N. F., Echenique, J. I. R., & Marx, C. J. (2012). Clarity an Open-Source Manager for Laboratory Automation. *Journal of Laboratory Automation*.

Chapter 2

Development of an optimized medium, strain and high-throughput culturing methods for *Methylobacterium extorquens*.

A description of the medium and strains developed to allow the experiments to be performed.

This work was done in collaboration with Maria E. Kaczmarek, Lewis. M. Ward and Paige K. Swanson.

Abstract

Methylobacterium extorquens strains are the best-studied methylotrophic model system, and their metabolism of single carbon compounds has been studied for over 50 years. Here we develop a new system for high-throughput batch culture of *M. extorquens* in microtiter plates by jointly optimizing the properties of the organism, the growth media and the culturing system. After removing the cellulose operon in *M. extorquens* strains AM1 and PA1 to prevent biofilm formation, we found that currently available lab automation equipment, integrated and managed by open source software, makes possible reliable estimates of the exponential growth rate. Using this system, we developed an optimized growth medium for *M. extorquens* using response surface methodologies. We found that media that used EDTA as a metal chelator inhibit growth and led to inconsistent culture conditions. In contrast, the new medium we developed with a PIPES buffer and metals chelated by citrate allowed for fast and more consistent growth rates. This new *Methylobacterium* PIPES ('MP') medium was also robust to large deviations in its component ingredients which avoided batch effects from experiments that used media prepared at different times. It also allowed for faster and more consistent growth than other media used for *M. extorquens*. This combination of new strains, medium and measurement system allow for the simultaneous measurement of the growth rate of 1,920 cultures with single observations having a precision within 2%.

The α-proteobacterium *Methylobacterium extorquens* has served for over 50 years[1] as the premier model system for uncovering the genetic basis of growth on C_1 compounds (i.e., methylotrophy) [2]. A significant number of genetic tools have been developed for *M. extorquens* over the past decade [3,4,5,6,7], and complete genome sequences are now available

for four strains of *M. extorquens*, as well as four genomes of other species in the genus [8,9]. This combination of genetic tools and genomic information has catalyzed experimental and computational analysis of systems-level properties [10,11], as well as the expansion of research into topics ranging from the natural ecology of *Methylobacterium* as a leaf epiphyte [12] to the adaptation of *M. extorquens* strains in the laboratory [13,14,15].

To facilitate quantitative analysis of the physiology of *M. extorquens*, we sought to develop a system for high-throughput growth rate assays. Studies using high-throughput methods to analyze growth rates have typically grown facultatively anaerobic organisms such as *Saccharomyces cerevisiae* or *Escherichia coli* in microtiter plate cell culture systems (see for example [16,17]). However, as *M. extorquens* are strict aerobes, this makes growth in traditional microtiter plate systems challenging, because oxygen transfer can be poor and stratification can readily occur [18,19]. An earlier attempt to grow *M. extorquens* AM1 strains in 96-well plates [15] resulted in informative quantitative patterns between strains, but the growth rate was extremely inconsistent through time and slower than what had been measured in flasks [14]. We therefore asked whether a combination of currently available products could be used or modified to allow for high throughput measurement of cultures of *Methylobacterium*.

Broadly speaking, automatic growth curve systems can be divided into two groups [19]. The first type of system is completely integrated where an incubator, shaker and plate reader are all contained in one physical instrument. Although such systems are typically easy to employ because they do not require integration of different instruments, they usually have limited capacity and can read and incubate only 1-2 plates at a time, and thus ~100-200 cultures

simultaneously. The second type of systems are those which have a physically separate shaking incubator where many plates can be grown, and a robotic arm which periodically moves plates from the incubator to a plate reader for measurements. A system like this was used in the robotic scientist project used to study *S. cerevisiae* [20], and has also been successfully employed in the measurement of growth rates of *E. coli* [17]. Because these systems allow for the measurement of several plates simultaneously, we used a collection of instruments with a separate incubator, plate reader and robotic arm as the starting point for this research.

High throughput growth rate assays not only require a robotic system that can take the necessary measurements, but also require culture conditions and a medium that allows for robust, repeatable growth rate measurements. A well-designed medium should not introduce biases or batch effects as it is remade or different reagent stocks are used. The medium formulation should also maintain a reasonably consistent environment throughout the growth cycle, and appropriately buffer any pH changes due to excretion or consumption during culture growth. At the outset of this study, it quickly became clear that the medium our lab had historically used, which is a variant of 'Hypho' medium (Table 2.1), did not fit these criteria. This medium uses phosphates as the pH buffer and contains an EDTA-chelated trace metal mix.

Table 2.1: Old Medium Recipe

Concentrations in Final Medium

Chemical	Concentration	Purpose
K_2HPO_4	14.5 mM	Buffer/Nutrient
NaH_2PO_4	18.8 mM	Buffer/Nutrient
$(NH_4)_2SO_4$	3.8 mM	Nutrient
$MgSO_4$	0.8 mM	Nutrient

General Comments

This medium uses phosphates as the buffer and the pH of the final media is determined by the relative concentration of the monobasic and dibasic phosphate components. A media equivalent to this one can also be made by switching the cation used in the phosphates as long as the relative concentration stays the same (such that $K_2HPO_4 \rightarrow Na_2HPO_4$ and $NaH_2PO_4 \rightarrow KH_2PO_4$). This simple recipe does not include Calcium or trace metals. Historically, we have added the 1000X "Vishniac" trace metal mix shown below. We write "Vishniac" in quotes because although it is similar to a recipe originally presented by Wolf Vishniac and Melvin Santer, it is not identical.

Preparation - This media can be prepared by combining two stock solutions.

Recipe (for 1 L): 100 mL 10X P-solution
100 mL 10X S-solution
800 mL deionized H2O

P-solution (10x): K_2HPO_4 25.3g (33.1 g K_2HPO_4 * 3 H_2O)
NaH_2PO_4 22.5g (25.9g NaH_2PO_4* H_2O)
in 1L of deionized H_2O

S-solution (10x): $(NH_4)_2SO_4$ 5g
$MgSO_4$ * 7 H_2O 2g (0.98 g $MgSO_4$)
in 1L of deionized H_2O

We experienced multiple problems with this medium, at least some of which were shown to be caused by metal limitation. In some cases, freshly prepared medium did not support growth at all until the medium had aged for several days and as a result different batches of media would produce distinctly different growth rates. We also found that measured growth rates differed depending on if the culture was being grown in plastic microtiter plates or in glassware. Further, during a long-term evolution experiment using this old medium recipe, cultures evolved to overcome trace metal limitations. One beneficial mutation that appeared in an experiment increased expression of a metal transporter [21]. This mutation specifically increased cobalt uptake, which is necessary for synthesis of the vitamin B_{12} that *M. extorquens* AM1 uses in a glyoxylate-regeneration pathway [21]. An independent analysis based upon quantitative metabolomics came to the same conclusion about this medium [22].

As cobalt and other trace metals are present at quantities typical of other media, we suspected the problems we observed were largely caused by the use of EDTA to chelate the metal cations in the media. EDTA has long been known to inhibit the growth of *M. extorquens* on methanol [23,24,25], and may sequester the metal cations so they are inaccessible to the cells. Consistent with this hypothesis, treating the media with light to degrade some of the EDTA allowed for improved growth [21]. However, we had not previously noticed these effects because robust growth could often be observed after the medium had aged by just a few days, the cells were less sensitive to the age of the medium when grown in glass, and quantitative differences were hard to detect without a robotic system. Collectively, it became clear that our traditionally used medium could not robustly support growth.

We therefore sought to create a new medium and ensure that no components in it were limiting or inhibiting growth. We used response surface methodologies [26] to design an optimal medium formulation and show that it was robust to deviations in the component concentrations. While designing a new media, in addition to varying the concentrations of the components in the original media, we also wanted to evaluate the use of alternative pH buffers as well as the effect of the total buffer concentration. Phosphate pH buffers are commonly used in media. They are easy to make and allow for the pH of a medium to be simply changed by varying the proportion of the dibasic and monobasic phosphate salts [27]. However, phosphate buffers have previously been found to inhibit the growth of some microorganisms [28], including *Methylobacterium* [29]. The effects of an alternative buffer such as PIPES, which performs well for Enteric bacteria [30], were unknown. Additionally, although past medium formulations for *Methylobacterium* have either used the strong chelator EDTA [14] or no strong chelator [31] (though phosphate buffers can act as a weaker chelator), we wanted to evaluate the effect of the alternative chelators NTA and citrate.

Here we report the results of these studies and a new *Methylobacterium* PIPES ('MP') medium that contains a citrate-chelated, trace metal solution with seven metals ('C7'). Furthermore, we present is the creation of genetically modified *M. extorquens* strains that do not form clumps while growing, and thus allow for substantially more consistent measurements of a culture's growth rate. With this combination of new media and strains, cultures growing in our 48-well microtiter plate culture system can reliably begin and maintain a phase of exponential growth,

allowing for growth rates to be measured with a mean squared error within 2% per replicate across hundreds, or even thousands, of cultures simultaneously.

Methods and Results

General Comments on the Methods – This section describes a series of experiments designed to create an optimized medium and experimental protocol to reproducibly measure the growth rates of *M. extorquens* strains. The two strains used in this study were *M. extorquens* PA1 (henceforth 'PA1'), which grows well on succinate but cannot double on methylamine in less than 10 hours, and *M. extorquens* AM1 (henceforth 'AM1'), which grows quickly on both succinate and methylamine. When the experiments tested AM1 on both succinate and methylamine and PA1 on succinate, we simply write that "all strains were tested on all substrates."

Growth rate measurements were performed using a robotic system composed of a shaking incubator that holds multi-well plates (Liconic USA LTX44 with custom fabricated cassettes) and a series of robotic instruments that can move these plates at regular intervals to a Perkin-Elmer Victor2 plate reader for optical density (OD_{600}, simply 'OD' hereafter) readings. A video showing the system and how the plates are moved is available at http://www.evolvedmicrobe.com/LabAutomation.html. The instruments were integrated and managed by Clarity, a recently described open source software program [32]. The entire robotic system is contained in a temperature- and humidity-controlled room set to 30 °C and 80% relative humidity. The OD readings output from this setup were parsed and fitted to an exponential model of cell growth using the methods described for the data analysis software reported in an accompanying companion paper [33].

Although the methods used for each of the experiments described here are presented independently, because optimization is an inherently iterative process the linear order of the experiments presented, their main question and simplified conclusions are outlined in Table 2.2 to unify them for the reader.

Table 2.2 – Summary of experiments and findings.

Experiment Number	Question	Answer
1	Can a system used for *E. coli* work well for growth rate measurements of *M. extorquens*?	No, 96 well plates do not shake adequately and the *M. extorquens* strains form clumps during growth.
2	Can changes be made to solve these problems?	Yes. 48-well plates have better mixing and deleting the Cel operon creates strains that do not form clumps in batch culture.
3	What chelators work well for *M. extorquens* media?	EDTA and NTA do not work well, but using citrate or no chelator does.
4	Are phosphate salts or PIPES a better pH buffer?	At higher buffer concentrations (48 mM) the phosphate salts are distinctly worse. At lower concentrations (30 mM) PIPES is still slightly better.
5	Does the sterilization method used for the citrate chelated C7 metal solution have an effect on the measured growth rates?	No, equivalent growth is seen when the solution is autoclaved, filtered or used without sterilization.
6	Can changing the concentrations of any of the medium components lead to a better medium?	No, the MP recipe appears optimal.
7	How does the new MP media compare to some other media used for *M. extorquens*?	MP media was the best one tested.

The medium our lab has historically used, henceforth the Hypho medium, was the starting point for our optimization process and for reference is given in table 2.1. The C7 metal mix we use in these experiments along with the complete MP medium recipe that came from this work is given

in table 2.3. When an experiment varied the levels of the C7 solution, to ease interpretation we described any alternate concentration as a multiple of the concentration in the final recipe. All media were tested at a pH of 6.75. The substrate concentrations varied slightly over the course of this research as we made small adjustments (15 or 17 mM methylamine, and either 5 or 5.6 mM succinate was used). We selected concentrations that allowed the cells to grow up to an equivalent maximum OD on both substrates, and that were as high as possible without observing a noticeable decline in the growth rate. When statistical models are employed, they are described using the standard formulas for the R/S languages (described online or at pg. 329 in [2]).

Table 2.3: Recipe for MP Medium

MP Media Recipe			FOR STOCK SOLUTIONS:		FOR 1 L MEDIA:	
		Molecular Weight (g)	CONCENTRATION	ADD TO 1 L dH_2O	FINAL CONCENTRATION	ADD
PIPES	$C_8H_8N_2O_6S_2$	302.37	300 mM (10X)	90.711 g	30 mM	100 mL
P solution			(100X)			
	$K_2HPO_4 \cdot 3\ H_2O$	228.22	145.0 mM	33.1 g	1.45 mM	10 mL
	$NaH_2PO_4 \cdot H_2O$	137.99	187.69 mM	25.9 g	1.88 mM	
$MgCl_2$	$MgCl_2 \cdot 6\ H2O$	203.3	2 M (4000X)	406.6 g	0.5 mM	250 µL
$(NH4)_2SO_4$	$(NH4)_2SO_4$	132.14	2 M (250X)	264.28 g	8 mM	4 mL
$CaCl_2$	$CaCl_2 \cdot 2\ H_2O$	147.02	2 M (100,000X)	294.04 g	20 µM	10 µL
C7-Metals	sodium citrate		(1000X)	ADD TO 500 mL dH_2O		
MIX IN ORDER LISTED ↓	($Na_3C_6H_5O_7 \cdot 2\ H_2O$)	294.1	45.53 mM	6705.5 mg	45.6 µM	1 mL
	$ZnSO_4 \cdot 7\ H_2O$	287.54	1.2 mM	172.52 mg	1.2 µM	
	$MnCl_2 \cdot 4\ H_2O$	197.9	1.0 mM	99 mg	1 µM	
	$FeSO_4 \cdot 7\ H_2O$	278.01	18 mM	2502 mg	18 µM	
	$(NH_4)_6Mo_7O_{24} \cdot 4\ H_2O$	1235.86	2 mM	1235.6 mg	2 µM	
	$CuSO_4 \cdot 5\ H_2O$	249.68	1 mM	124.8 mg	1 µM	
	$CoCl_2 \cdot 6\ H_2O$	237.93	2 mM	237.9 mg	2 µM	
	$Na_2WO_4 \cdot 2\ H_2O$	329.85	0.33 mM	54.4 mg	0.33 µM	
milliQ-H_2O						885 mL

Directions to prepare medium:

1. Make PIPES according to the preceding table and adjust to pH 6.75 by adding KOH.
2. Make additional stock solutions for P, $MgCl_2$, $(NH_4)_2SO_4$. Autoclave all.
3. To make stock solution for C7 METALS, mix metals in order listed, dissolving sodium citrate first. Dissolve each metal before adding the subsequent one. Autoclave.
4. To make media, mix all components together except $CaCl_2$. Autoclave. Add $CaCl_2$. ($CaCl_2$ is added afterwards to avoid calcium phosphates from being formed in the autoclave).

1. Initial tests and use of 48-well plates – We first tested whether we could reliably measure *M. extorquens* exponential growth rates using a robotic measurement system under similar conditions to those used by previous investigators with *E. coli* [3]. AM1 cultures were grown in a Microtest 96-well tissue culture treated plate (Falcon-35-3072). Cells were grown in buffered medium prepared using 14.5 mM of K_2HPO_4, 18.8 mM of NaH_2PO_4, 8 mM ammonium sulfate, 20 μM calcium chloride and the C7 metal mix that was left unchelated by not adding citrate with 17 mM methylamine·HCL added to the base medium. The mixture was aliquoted in 160 μL portions into wells of a 96-well plate. The growth curves from the initial tests in 96 well plates showed huge deviations from the exponential model, and were exceptionally noisy. We concluded that 96-well plates were inadequate for sustained exponential growth of *M. extorquens*. We re-evaluated this conclusion at the end of this project, after optimizing our strains and media, by again growing *M. extorquens* in MP media and in 96-well plates and still found the model of exponential growth to be strongly violated, as the cultures never achieved a steady growth rate and again showed poor growth characteristics (Fig 2.1).

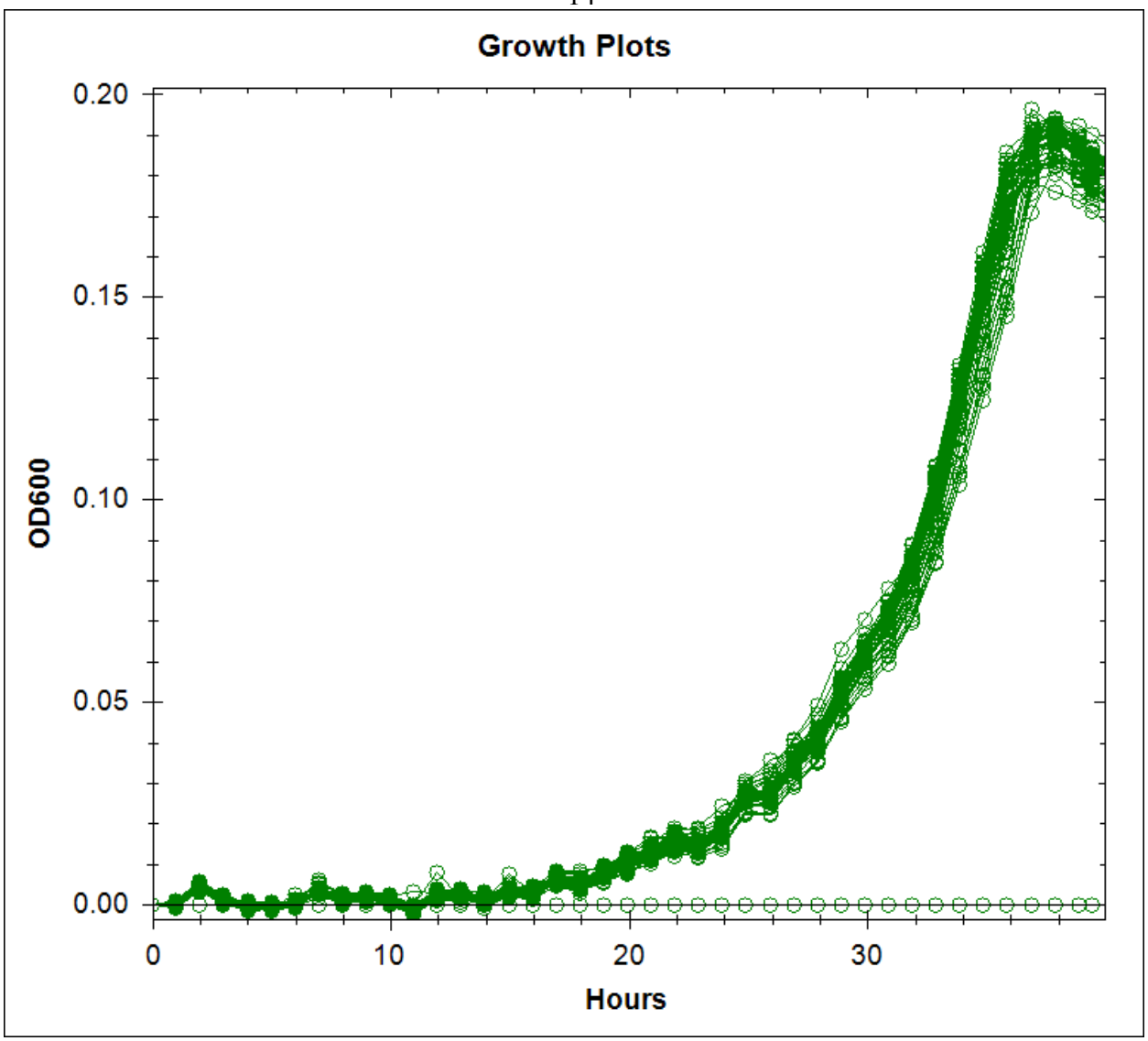


Figure 2.1 – OD through time for cells grown in a 96-well plate. The OD readings show a strange bump after approximately 30 hours, which occurred due to cells sedimenting in the well.

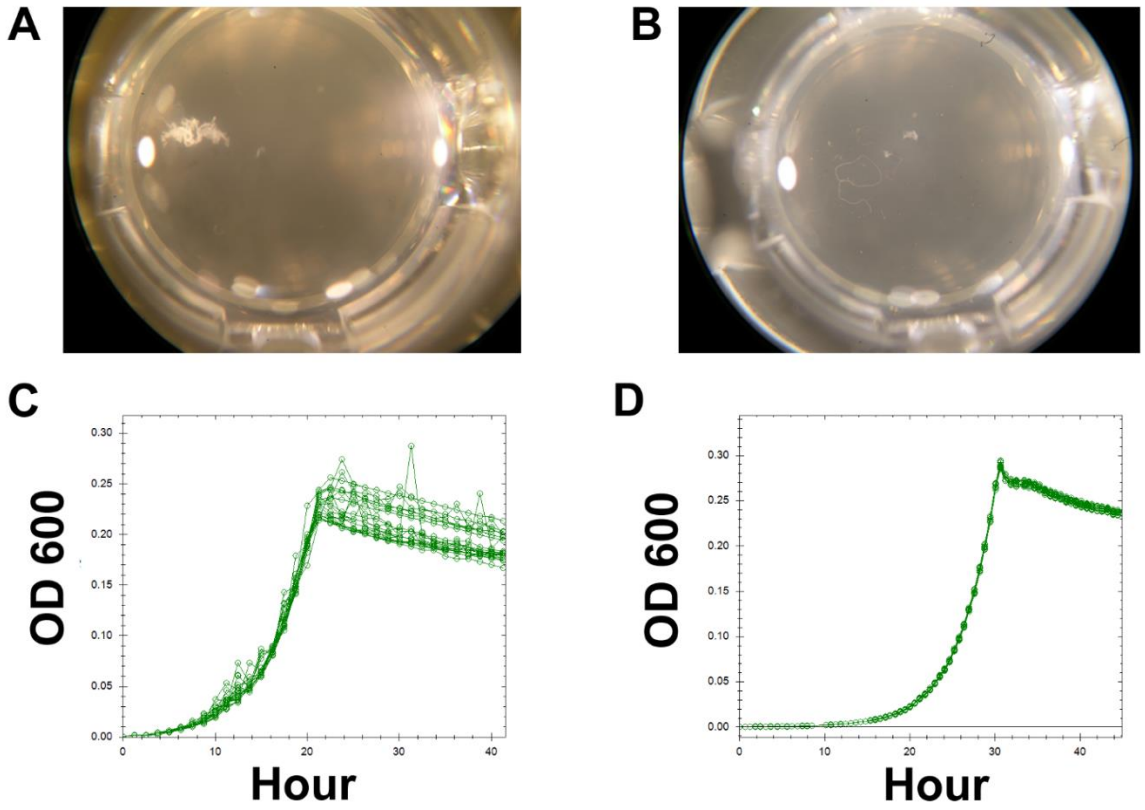


Figure 2.2 – A comparison of the growth characteristics of *M. extorquens* with (A) and without (B) the cellulose operon removed. The top row shows pictures of individual wells after growth. A large clump in the wild-type strain is indicated by a light blue arrow. Such clumps were not found for AM1Δ*cel*, occasionally small strands such as those shown in B can appear. The bottom row (C and D) shows example growth curves obtained for 12 replicates with or without the cellulose locus. The OD readings of wild-type were noisy and irregular (C), whereas readings from AM1Δ*cel* are more regular (D).

In contrast, we found that 48-well plates did allow for adequate mixing and consistent exponential growth (Fig 2.2D). We therefore altered the robotic system by installing new custom built racks so that it could use 48-well plates (CoStar-3548) instead of 96-well plates. In contrast to the 96-well plates where the medium did not appear to move within the well, the media in the 48-well plates rhythmically swirled around. We also tested a second type of 48-well plate, from the Cellstar line made by Greiner Bio-One (Catalog #677 102). Surprisingly, although medium in the CoStar plates visibly swirled while shaking, the meniscus in the Cellstar plates, as in the 96-well plates, stayed at approximately the same level and did not appear to move; correspondingly, cultures grew much more poorly in Cellstar plates. For all future work in this paper, we grew the cultures in CoStar plates in 640 μL per well with the incubator shaking at 650 RPM, as growth and the swirling of the liquid appeared to be as good or better than the range of other values tested.

2. Creation of a cellulose deletion strain – We noticed that cultures growing in microtiter plates formed clumps that moved around inside the wells (Fig 2.2A) and contributed to noise in the OD readings. Other researchers had found that an AM1 strain with a transposon insertion into its cellulose operon did not clump as frequently or as severely as the wild-type (A. Stöver and M. Lidstrom, personal communication). To replicate this effect in our strains without introducing the markers associated with the transposon insertion, we removed the cellulose operon from both of these strains by excising a 7,183 bp region containing this operon. Three genes suggested to be involved in cellulose synthesis—*celA*, *celB* and *celC*—were removed, as well as a portion of a gene with an unknown function (these genes are numerically annotated in the reference AM1 genome [4] as META1_1167, META1_1168, META1_1169 and META1_1170). The above

described ‘Δ*cel*’ alleles were constructed by joining two PCR products flanking the region to be removed from AM1 or PA1 in the *sacB*-based allelic-exchange vector, pCM433 [5]. These two plasmids, pLW17 and pLW18, were then used to obtain unmarked Δ*cel* strains CM2720 and CM2730, respectively. For simplicity, hereafter we refer to these as AM1Δ*cel* and PA1Δcel. We confirmed the deletion in the Δ*cel* strains by sequencing. Growth dynamics on the robotic system were strikingly different between the Δ*cel* strains compared to the corresponding Cel^+ wild-type cultures (Figs 2.2C, 2.2D)**.** The Δ*cel* strains showed significantly more stable curves without the sharp spikes in OD measurements previously observed (Fig 2.2D), and thus from this point forward all optimization was performed with these two strains.

3. Testing different metal chelators – We previously found that when EDTA was used to chelate metals in our medium with either methanol or methylamine used as the substrate, cultures often grew slowly [21], and sometimes would not grow at all in 48-well plates. To test other chelators and quantify how they affected the growth of *M. extorquens*, we compared the growth of rates of both AM1 and PA1 Δ*cel* strains on methylamine, succinate and also methanol, using five different chelator treatments. We tested three chelators (EDTA, NTA, and citrate), as well as two unchelated metal mixes that were prepared either immediately before testing or several months prior to the experiment. Trace metal solutions with different chelator treatments were prepared by adding each chelator to a solution otherwise identical to the C7-metal mix solution (Table 2.3) except made by excluding citrate from the recipe so that this base solution did not already contain a chelator. Five concentrations of the chelator concentration (given as a percentage of total moles of chelator relative to total moles of metal ions in the solution) were tested for both EDTA and NTA: 5%, 25%, 50%, 100% and 200%. Citrate was restricted to three

levels: 50%, 100% and 150%. Excepting the trace metal treatments, all media tested were identical to a Hypho medium recipe made while excluding that medium's own trace metal solution (Table 2.1).

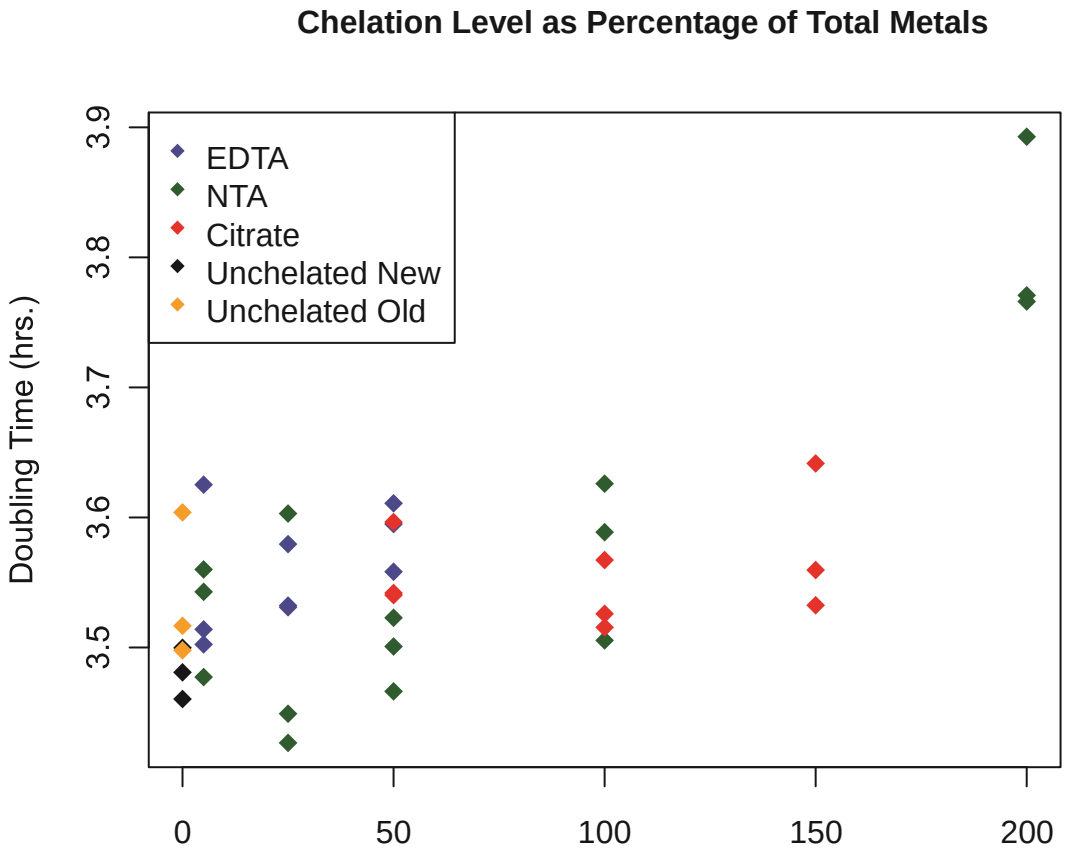


Figure 2.3 – Growth rates for AM1Δ*cel* growing on methylamine. The doubling times for the cells in the EDTA treatments at 100% and 200% of the trace metal concentrations are not shown as either those wells never had an observable increase in OD (the 200% treatment) or only did so after a very long time and never achieved a high enough OD to accurately measure the growth rate (the 100% treatment).

We found that neither EDTA nor NTA were acceptable chelator options, but that either citrate or no chelator were. On methylamine, except for the 100% and 200% EDTA treatments, AM1 was

able to grow on all chelated and unchelated medium (Fig 2.3). Cultures of the 200% EDTA treatment never grew and the 100% EDTA treatments began to only exhibit very slow growth after ~85 hours. These two treatments were therefore excluded from further analysis. We analyzed the remaining treatments using a linear model where growth rate was a function of the term for chelator treatment and an interaction between each chelator and the relative concentration of chelator. In agreement with the general pattern in Fig 2.3, which shows slower growth at higher NTA concentrations, the only statistically significant term in this model was the interaction between NTA and chelator concentration ($p = 0.01$), evincing that NTA reduces growth rate as its concentration increases (Fig 2.3).

On succinate the chelator utilized had less of an effect upon growth. The only meaningful difference was observed for the 200% EDTA treatment, which was much worse than the other treatments having a mean growth rate 6.7 units of the standard error for the other treatments away from the mean of all other treatments on AM1Δ*cel*, and 3.85 standard errors away from the mean of all other treatments on PA1Δ*cel*. We concluded from this experiment that unchelated or citrate chelated metals would not inhibit growth, and chose citrate chelated metals for our medium as they did not form precipitates as the unchelated ones did.

4. **PIPES vs. phosphates as a pH buffer -** We tested the effect of varying the type of pH buffer used and its concentration on growth rate for all strains (AM1Δ*cel* and PA1Δ*cel*) on all substrates. We ran a full factorial experiment using two buffers, PIPES or phosphates, at two concentrations, either 30mM or 48mM. The four possible treatments were added separately to

the following additional medium ingredients: 5 mM ammonium sulfate, C7 metals and 3.33 mM P-Solution (Table 2.1), which was added as a phosphate source.

The growth rates obtained from each replicate in this experiment are displayed in Figure 2.4. We analyzed the data separately for each strain and substrate combination using a full linear model with effects for the Buffer and Buffer:Concentration interactions. For each strain/substrate combination, all four treatment groups were significantly different (p-values< 0.01) except for PA1Δ*cel* growing with PIPES on succinate, where the growth rate was not significantly different at either 30 mM or 48 mM ($p = 0.82$). Although PIPES medium at 30 mM had the highest estimated growth rate, the differences were slight, with an estimated effect of 1%-5% of the growth rate depending on the substrate. At 48 mM concentrations, however, the differences between PIPES and phosphates were far more pronounced, with the phosphate buffer causing substantially slower growth (Figure 2.4). Based on these results, we selected PIPES as our medium's pH buffer.

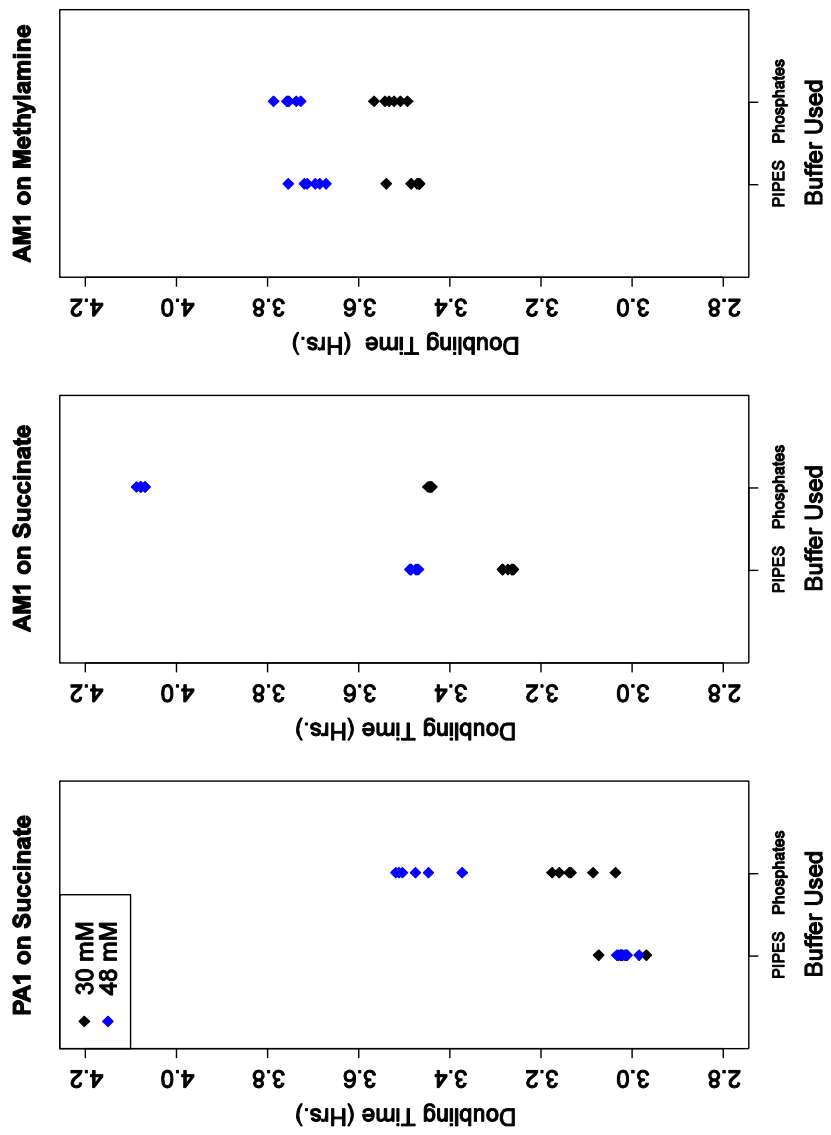


Figure 2.4 – Comparison of the growth rates obtained for all strains and substrates with either PIPES or Phosphates used as the buffer and at two different concentrations. Red symbols indicate 48 mM buffer concentration treatments and black symbols indicate 30 mM buffer concentrations.

5. Effect of sterilization protocol on C7 Metal Solution – We tested whether the method used to prepare the C7 metal solution would affect the growth rate. After making our standard C7 metal mix, we either autoclaved, filter-sterilized, or left it untreated before adding it to the base medium. We grew AM1Δ*cel* on methylamine (the most metal sensitive growth condition) and found no significant differences among the growth rates measured in these three treatments ($F_{2,19}$ = .18, p = 0.84). We concluded that metal preparation did not greatly affect growth rate.

5. Optimizing concentrations of medium components - Having selected PIPES as the buffer and the citrate chelated C7 solution as the source of trace metals in the new medium, we next tried to ensure that the new medium would be robust to reasonable variations in the exact concentration of all its component solutions (the $MgCl_2$, the phosphates, the C7 metal mix, and the $(NH_4)_2SO_4$) and we also wanted to look for any further optimizations that might be possible. To do this, we used a full factorial design with a central point [26], which is an experimental design that allows one to not only estimate the effect of individually altering the concentration of any component, but also the effect of simultaneously altering several components. Implementing this design requires one to pick a range of concentrations to test for each component, and we selected for each a range of variation that far exceeds likely differences between replicated preparations of the medium. The 3 levels selected for each component were as follows: $(NH_4)_2SO_4$ –3 mM, 5 mM and 7mM; $MgCl_2$ – 0.125 mM, 0.25 mM and 0.5 mM; C7 metal mix – 0.5X, 1X and 1.5X; and phosphates – 2.66 mM, 3.33 mM and 4 mM. Each set of media components was made independently in 3 different batches and replicated 3 times (n = 153 for each strain/treatment) for both strains on both substrates. In addition, because each plate can

hold only 48 samples, the plate/incubator-position was included as a blocking covariate in the analysis (henceforth referred to as the "Slot" factor).

We found that growth was very robust to the large changes in the concentrations of the different components. For every strain and substrate tested, the data were consistent across different treatments. The estimated standard error in a linear model fit to the specific treatment and blocking plate factor (i.e. the pure error) was between 1% and 2% of the estimated growth rate at the central point. Using the same model, the greatest difference in estimated growth rates (defined as the growth rate of the fastest estimated treatment effect subtracted from the slowest estimated treatment effect) was only 4%, 6% and 9% for AM1Δ*cel* on methylamine, AM1Δ*cel* on succinate and PA1Δ*cel* on succinate, respectively.

To determine which medium components contributed to what little variation was observed across treatments, as well as to discover any further optimizations, we analyzed data from each strain and substrate combination using linear models with the concentration of each media component as a factor coded using a -1, 0, +1 scheme. Full models with complete interactions for all components as well as the slot blocking variable were initially used to fit the data. To ensure that our conclusions were robust to model-selection criteria we ensured that the conclusions held over a variety of models that could be reduced from the complete model using a combination of approaches including the Akaike Information Criteria (AIC) criteria and the effect-hierarchy principle. We also examined partial residual plots to evaluate any need for quadratic terms in the model.

For all strain and substrate combinations, we found that the concentrations of $(NH_4)_2SO_4$ and $MgCl_2$ in a medium were the only two significant factors, though their estimated effects were very small. Increasing the $MgCl_2$ relative to the lowest value appeared to be mildly beneficial for all strains on all substrates. The estimated effect was statistically significant though less than 0.5% of the growth rate for the central point treatment when AM1Δ*cel* is grown on methylamine ($MgCl_2$ =.001, p = .0001 in model: GrowthRate ~ Slot + $MgCl_2$ + $(NH_4)_2SO_4$) and the effect was less than 1% of the central treatment's growth rate when grown on succinate ($MgCl_2$ =.0024, p = 3.13e-6 in model: GrowthRate ~ Slot + $MgCl_2$ + $(NH_4)_2SO_4$). For PA1Δ*cel* grown on succinate, the main effect of increased $MgCl_2$ concentration was only 2% of the central point's growth rate, though this interpretation is slightly complicated by an interaction term in the model that slightly reduces the total beneficial effect when the ammonium concentration is also increased ($MgCl_2$ =.0055, p < 2e-16 in model: GrowthRate ~ Slot + $MgCl_2$ + $(NH_4)_2SO_4$ + $MgCl_2$: $(NH_4)_2SO_4$).

As increased $MgCl_2$ resulted in faster growth, we sought to optimize its concentration with an additional experiment that used a range of higher concentrations. We grew all strains and substrates in media with $MgCl_2$ concentrations of 0.25, 0.5, 0.75, 1.5 or 2 mM, with all other media components set to the midpoint of their range in the previous experiment. Although we found no significant effect in this later experiment for the concentration of $MgCl_2$ on PA1Δ*cel* ($F_{1,39}$ = .84, p = 0.37), there was a very slight negative effect of increasing $MgCl_2$ for growth of AM1Δ*cel* on either substrate (p-values in both regressions < 0.05, with both estimated effect sizes less than 2% of the mean growth rate per mM increase in $MgCl_2$); however, this negative effect was largely due to a decrease in the growth rate that began at the 1.5 or 2mM levels and no significant differences were found when comparing the mean growth rates over the 0.25-0.75

mM levels for any strain or substrate concentration (p-values > 0.5). Our result that $MgCl_2$ could limit growth at concentrations at or below 0.125 mM is consistent with a previous study that found $MgCl_2$ limited the growth of AM1 on methanol at concentrations below 0.121 mM [34]. For this reason, we set the $MgCl_2$ level at 0.5 mM in MP medium, and kept the concentration of all other medium components at the midpoint of their tested levels.

In contrast to the consistently beneficial effect of increased $MgCl_2$, the effect of increasing the $(NH_4)_2SO_4$ concentration was very slightly beneficial on methylamine and very slightly deleterious on succinate. Effect sizes using the previously specified models were as follows: succinate AM1Δ*cel* = 0.001, PA1Δ*cel* = 0.004; methylamine: AM1Δ*cel* = -0.0009; all 3 p-values for each estimated effect < 0.0006). As the effect size in all cases was less than 0.4% of the mean growth rate for each strain and substrate, and because the direction of the effect depended on the substrate, we did not further consider this variable for optimization. One possible explanation for this divergent result is that ammonia is liberated during methylamine consumption, and thus additional nitrogen in MP medium is unnecessary for growth on this substrate.

6. Comparison to other media – To validate that our medium, henceforth "MP", compared well to other formulations currently used to grow *M.extorquens*, growth rate was compared for AM1Δ*cel* growing on methylamine and methanol, as well as AM1Δ*cel* and PA1Δ*cel* on succinate, in MP and four other media. The first medium we tested was our historically used variant-Hypho (aged for over four weeks). The second and third media tested were phosphate-buffered media that differed in initial pH (second media: initial pH = 6.7 for growth on multi-

carbon compounds, henceforth “Phosphates-multi-C”; third media: initial pH = 7.1 for growth on C_1 compounds, here “Phosphates-C_1” [35]). The rationale behind testing different pH levels was to partially counter the tendency that growth on multi-C substances increases pH, whereas the opposite is seen for C_1 compounds. The final medium we compared, Choi medium [36], is a *Methylobacterium* medium developed to aid poly-β-hyroxybutyric acid (PHB) production and has an exceptionally metal-rich formulation; total trace metals are in the mM range instead of the μM range. A comparison of the concentrations of main components of each of these media is given in Table 2.4.

Table 2.4 – A comparison of the main components of the different medium formulations compared.

	Old Media (Hypho-Variant)	Phosphates C_1	Phosphate Multi-C	MP	Choi
Buffer	Phosphates	Phosphates	Phosphates	PIPES	Phosphates
pH	6.7	7.1	6.7	6.7	6.8
Buffer Conc.	33.3 mM	20.7 mM	20.7 mM	30 mM	24.6 mM
Chelator	EDTA	EDTA	EDTA	Citrate	None
Calcium	9.98 uM	20.41 uM	20.41 uM	9.98 uM	13.6 mM
Total Metals (excluding Ca and Mg)	12.66 uM	63.16 uM	63.16 uM	12.66 uM	13 mM
Nitrogen	1.89 mM	30.29 mM	30.29 mM	4 mM	5.68 mM
Phosphates	33.3 mM	20.7 mM	20.7 mM	3.33 mM	24.6 mM
Magnesium	0.81 mM	0.81 mM	0.81 mM	0.5 mM	1.83 mM

On C_1 compounds, strains grown in MP medium grew faster than in all other media (Fig. 2.5). With methylamine as the substrate, the growth rate on MP was estimated to be 11% faster than on our older variant-Hypho, and 15% faster than on Phosphates-C_1 medium (all p-values < $1x10^{-6}$). With methanol as the substrate, due to evaporation, the cultures did not achieve an OD over 0.1 and could not be fit over the same range of OD values; however, when fit over an OD range of 0.01-0.07, the MP medium was estimated to be 7% and 17% faster than on variant-Hypho and Phosphates-C_1 (p-values < 1e-6), respectively. Unfortunately, no comparisons could be made to the Choi media as it produced data that was too noisy for meaningful analysis. Although the Choi medium did appear to have growth rates similar to the other media tested, the large concentration of unchelated metals in Choi medium formed dense precipitates on the bottom of the wells, making it difficult to set a well's blank values and causing highly erratic OD measurements throughout the growth period. For this reason meaningful quantitative comparisons could not be made and we concluded that Choi medium could not be used for growth rate measurements in microtiter plates.

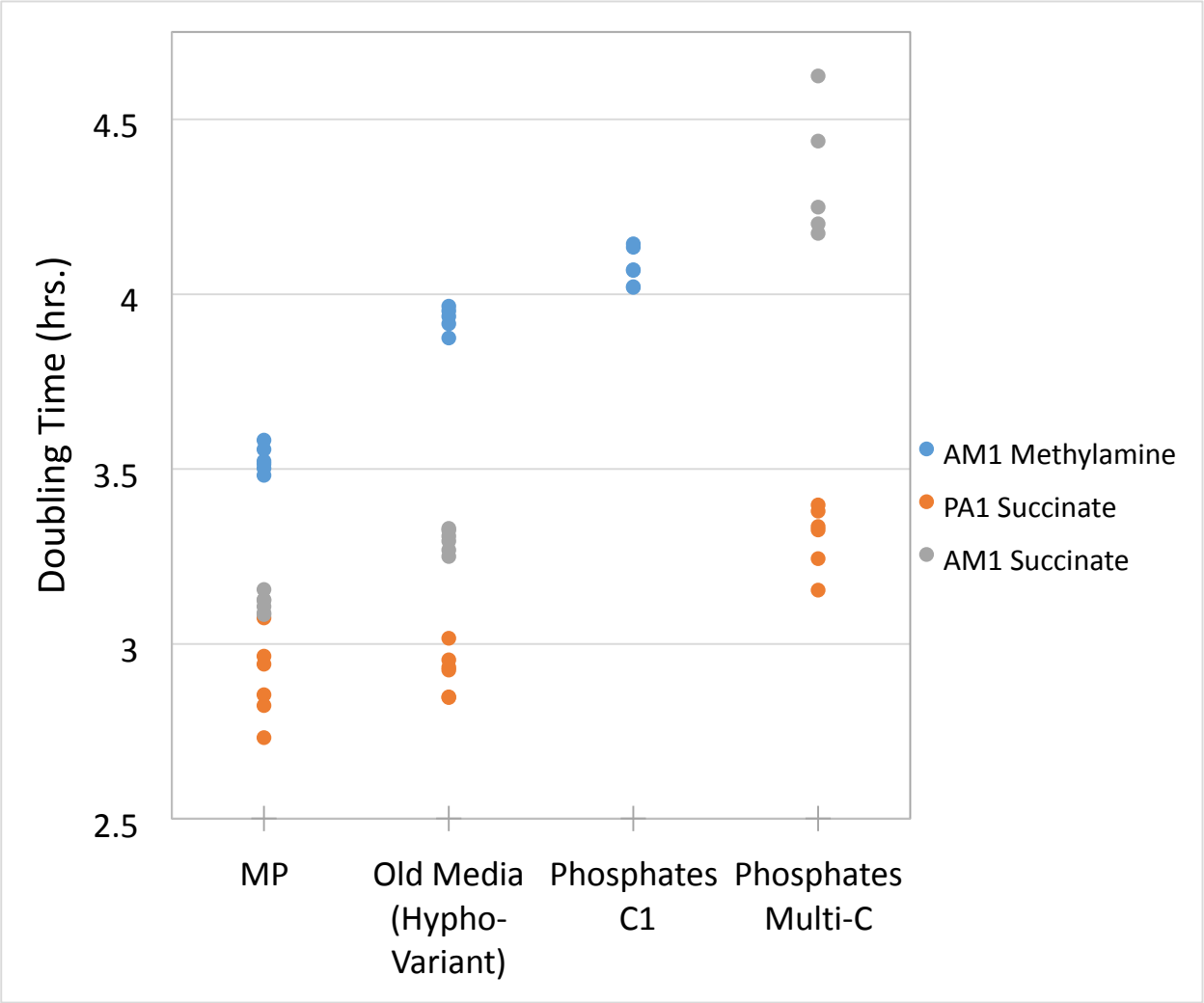


Figure 2.5 – Growth rates of AM1Δ*cel* and PA1Δ*cel* on different media formulations and carbon sources. The data for the Choi medium is not shown because the large amount of metal precipitates in this media introduced too much noise into the OD readings so that accurate growth rate estimates could not be made.

On succinate, strains grown in MP medium performed as well as or better than the other media we tested (Fig. 3, Fig. S3**)**. For AM1Δ*cel* growing on succinate, the mean growth rate was

estimated to be 6% faster with MP ($p = 1.74x10^{-8}$), while for PA1Δ*cel* there was a smaller but still significant improvement of 1.7% compared to Hypho medium ($p = .01$). Although the Phosphate-multi-C medium initially appeared to give roughly equivalent growth rates to variant-Hypho or MP. However, its growth rate noticeably slowed as OD increased in a clear violation of the exponential growth model (Fig. S3).

Discussion

We have developed a high throughput system to accurately and reproducibly measure growth rates of *M. extorquens* strains by integrating robotic instruments, genetically modifying strains and designing a new growth medium. Our system is capable of simultaneously measuring the growth rate of 1,920 cultures of *M. extorquens* if each 48-well plate is measurement every 50 minutes, and the standard error on each inferred growth rate is estimated to be less than 2% of the actual value. The deletion of the *cel* operon in both *M. extorquens* strains was a large step towards being able to make such accurate measurements. Even with otherwise optimal growth conditions, the formation of clumps of grouped cells (Fig 2.2A) that still contained this operon obscured the relationship between increasing OD measurements and the total increase in biomass, and led to data that was too noisy to allow for precise growth rate measurements.

The use of 48-well plates instead of 96-well plates was another important factor and is a substantial difference between our system and others. It was also the only change that required us to custom fabricate components for our robotic system rather than simply combine available products; the slots that hold plates in the incubator tower had to be redesigned to fit the larger plates. Although a previous study also found that 48-well plates provide significantly better mixing [37], we were very surprised that the two types of nearly identical 48-well plates we

tested had such drastically different mixing and growth characteristics. These two plate types have indistinguishable standard dimensions and are both made of polystyrene. However, the CoStar plates are tissue culture treated, while the Greiner plates are not, and we suspect this explains the difference in how the medium swirled in their wells and whether they allowed for stable growth rates.

The new medium formulation optimized for *M. extorquens*, MP, overcomes inconsistencies in other media and is robust to minor variations in its components. There are several aspects of this media that make it robust relative to other media. A major difference between MP and other media commonly used for *M. extorquens* is the decision to use citrate instead of EDTA as a chelator. Although EDTA had clear disadvantages and did not allow for consistent measurements, because a citrate chelator could be a possible carbon source for some organisms it is sometimes avoided. Notably, citrate is only present in MP at a concentration of 45 μM, which is 100-fold or more below substrate concentrations utilized for growth [38]. However, our results indicate that using citrate does not affect growth dynamics; when the total concentration of the C7 solution was varied as part of our experiments no differences were found. We also found citrate preferable to not using any chelator. Although cultures appeared to grow as fast on an unchelated version of the C7 metal mix (made by simply excluding the citrate), the oxidation state of metal cations in a liquid solution can more readily change if they are not chelated and at equilibrium oxidized and unchelated metals may almost be entirely in biologically unavailable forms if they have largely precipitated out of solution [39]. The unchelated C7 metal mix appears susceptible to these problems as it does form a precipitate, making it difficult to ensure consistent concentrations in different aliquots, and it also changes color as it ages over several

months. Thus, in designing a new medium, we have chosen to use citrate as the chelator due to better optical properties and apparently greater stability.

The MP medium, similar to other media used for *Methylobacterium* species [36,40,41], is unusually metal rich. Many of the metals in it are present at concentrations above 1 µM, whereas most media for bacteria provides each trace metal at a concentration between 0.01 and 1 µM, as they are often toxic at higher concentrations [28]. The higher metal concentrations in MP medium are not toxic to *M. extorquens* however, as we found no advantage on either succinate or methylamine to increasing or decreasing the concentrations by 50%.

Interestingly, we found that growth on single carbon compounds requires higher metal concentrations than growth on multiple carbon compounds and this likely explains why growth on succinate is relatively unaffected by the type of chelator used. In preliminary studies leading up to the work presented here we found that when AM1 grows on succinate, the concentration of the C7 solution can be reduced to a small fraction of its level in MP medium without affecting the growth rate. In contrast, the metals must be maintained much closer to the unusually high concentrations in MP medium for growth on single carbon compounds. In particular, for growth on methylamine, copper, a component of the amicyanin protein thought to receive electrons from methylamine dehydrogenase [42] was found to be the first metal to limit growth in earlier tests. However, we did not find that all metals were required in measurable concentrations. As an example, the C7 metal solution, unlike the other media we compared, also contains tungsten as a component. We wanted to ensure that tungsten was available in adequate amounts as it has been shown to be used by a formate dehydrogenase enzyme in *M. extorquens* [43]. However, we

were unable to show either a positive or negative effect of explicitly adding tungsten to the medium, as equivalent growth rates are obtained with or without it, implying that some ambient source of tungsten is usually sufficient or the cofactor is unnecessary.

Although the final MP medium formulation for both AM1Δ*cel* and PA1Δ*cel* is robust to large deviations in the concentrations of almost all its components, the exception is a trade-off we detected when selecting the concentration of the pH buffer. Low buffer concentrations can create initially faster growth rates, while higher concentrations allow for a slower but more consistent growth rate over a larger range of OD values. In this study we only demonstrated that growth rate decreases with an increasing buffer concentration by comparing growth rates at concentrations of either 30 or 48 mM, but some preliminary work suggested that the growth rate of *M. extorquens* appeared to be slightly faster when the concentrations of the pH buffer was below 30 mM. However, it is difficult to take reliable growth rate measurements of cultures grown in concentrations below 30 mM as the growth rate noticeably declines as the culture grows, making it hard to measure any single consistent growth rate. This was also seen in the Phosphates-multi-C medium we tested, which uses a 20.7 mM concentration of buffer leading to a decreasing growth rate at higher OD values (Fig 2.6). We chose a 30 mM buffer concentration (with an additional 3.33 mM of buffering provided by the phosphate solution) for MP medium as a compromise that allowed stable measurements over a range of OD values the growth curve, but did not appear to significantly hinder growth relative to lower concentrations.

At this 30 mM buffer concentration, using PIPES instead of a phosphate buffer only slightly increases the growth rate. Although the two buffers behaved similarly, we chose the slightly

harder to prepare PIPES buffer because the phosphates had a tendency to occasionally form a small amount of white “snow” in the medium (presumably calcium phosphate precipitates). This snow not only changes the composition of the medium, but we felt that as some particles can be approximately the size of a cell, if cells were being counted by flow-cytometry the snow might also lead to false positive counts. Furthermore, at higher buffer concentrations PIPES is clearly superior to phosphates, particularly on succinate (Fig 2.5). If it was desired to grow *M. extorquens* at very high densities, one could alter the MP medium formula to increase the PIPES buffer concentration without seeing as substantial a decrease in the growth rate as one would if they used phosphates. This may be applicable to current media formulations designed to optimize the production of industrial products using *Methylobacterium* [40,41], and these media might benefit by switching the buffer from phosphates to PIPES. It is also possible that our strains with the deleted cellulose operon might make better candidates for industrial production strains, as presumably less biomass is being channeled towards production of extra-cellular carbon.

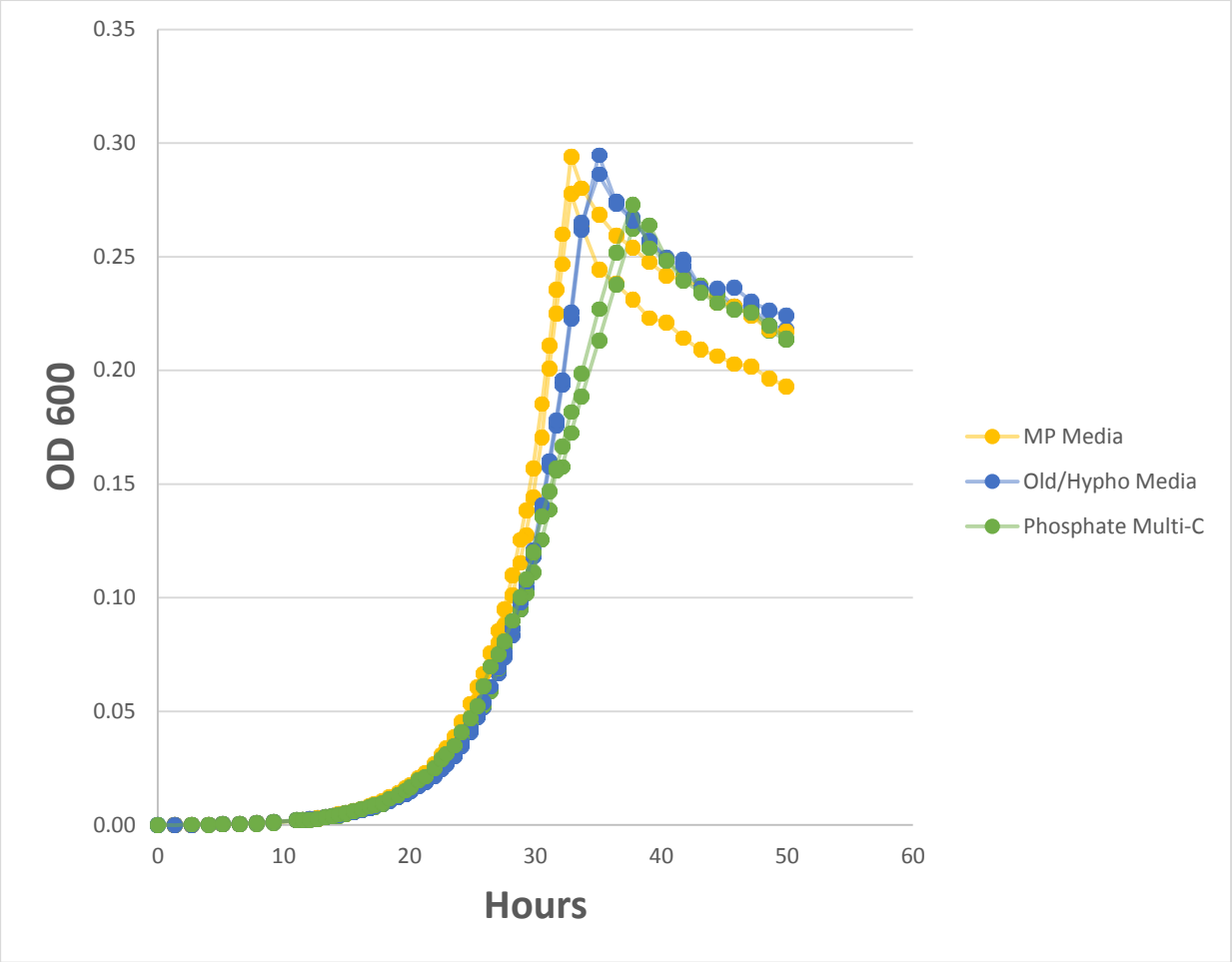


Figure 2.6 – OD through time plots of AM1Δ*cel* growing on succinate in three different media. Although the Phosphate multi-C medium initially appears roughly equivalent to the other two media, it shows a noticeable slow down once the OD reaches 0.15.

Our medium, strains and instrumentation allows for precise measurements of growth rates at a very large scale. As the study of the physiology of *M. extorquens* has become increasingly quantitative, the need to move beyond the "-, +/-, +, ++" categorization of microbial growth to more precise measurements of the growth rate has become ever more important. Interesting questions that can now be more effectively answered range from exploring the growth of strains

across a wide spectrum of continuously-varied enzyme levels via a regulated promoter [4] to exploring differences in the lag time required to switch between substrates [13]. Furthermore, from an evolutionary perspective small differences in the growth rate can be tremendously important. Beneficial mutations in populations that have evolved in batch culture, where growth rate is the primary selective component, can commonly be less than 10% and tend to decrease dramatically as adaptation proceeds. Being able to evolve strains in a consistent media environment and measure their growth rates as accurately as our system allows can therefore provide great insight into the adaptive dynamics of evolving populations. It is exciting that we finally have the ability to measure at scale the growth dynamics of the metal-hungry, strictly-aerobic and ever-interesting *M. extorquens*.

Acknowledgements

We would like to thank Tirthankar Dasgupta who many years ago introduced us to the beautiful efficiencies and many-at-once power of fractional factorial designs. Emily Kay and members of the Marx lab provided excellent feedback on our manuscript. N. F. D. was supported by an NSF graduate student fellowship and this research was funded by an NSF CAREER grant to C. J. M. (DEB-0845893).

References

1. Large P, Peel D, Quayle J (1961) Microbial growth on C1 compounds. 2. Synthesis of cell constituents by methanol-and formate-grown Pseudomonas AM1, and methanol-grown *Hyphomicrobium vulgare*. Biochemical Journal 81: 470.

2. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. J Bacteriol 185: 2980-2987.

3. Marx CJ (2008) Development of a broad-host-range sacB-based vector for unmarked allelic exchange. BMC Res Notes 1: 1.

4. Chou HH, Marx CJ (2012) Optimization of gene expression through divergent mutational paths. Cell reports.

5. Marx CJ, Lidstrom ME (2004) Development of an insertional expression vector system for Methylobacterium extorquens AM1 and generation of null mutants lacking mtdA and/or fch. Microbiology 150: 9-19.

6. Marx CJ, Lidstrom ME (2002) Broad-host-range cre-lox system for antibiotic marker recycling in gram-negative bacteria. BioTechniques 33: 1062-1067.

7. Marx CJ, Lidstrom ME (2001) Development of improved versatile broad-host-range vectors for use in methylotrophs and other Gram-negative bacteria. Microbiology 147: 2065-2075.

8. Marx CJ, Bringel F, Chistoserdova L, Moulin L, Farhan UHM, et al. (2012) Complete genome sequences of six strains of the genus methylobacterium. J Bacteriol 194: 4746.

9. Vuilleumier S, Chistoserdova L, Lee MC, Bringel F, Lajus A, et al. (2009) Methylobacterium genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. PLoS One 4: e5584.

10. Peyraud R, Schneider K, Kiefer P, Massou S, Vorholt JA, et al. (2011) Genome-scale reconstruction and system level investigation of the metabolic network of *Methylobacterium extorquens* AM1. BMC Syst Biol 5: 189.

11. Skovran E, Crowther GJ, Guo X, Yang S, Lidstrom ME (2010) A systems biology approach uncovers cellular strategies used by Methylobacterium extorquens AM1 during the switch from multi- to single-carbon growth. PLoS One 5: e14091.

12. Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, et al. (2012) Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. ISME J 6: 1378-1390.

13. Lee MC, Chou HH, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of Methylobacterium to distinct growth substrates. Evolution 63: 2816-2830.

14. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332: 1190.

15. Marx CJ (2012) Recovering from a bad start: rapid adaptation and tradeoffs to growth below a threshold density. BMC Evol Biol 12: 109.

16. Warringer J, Blomberg A (2003) Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae*. Yeast 20: 53-67.

17. Bollenbach T, Quan S, Chait R, Kishony R (2009) Nonoptimal microbial response to antibiotics underlies suppressive drug interactions. Cell 139: 707-718.

18. Hermann R, Lehmann M, Büchs J (2003) Characterization of gas–liquid mass transfer phenomena in microtiter plates. Biotechnol Bioeng 81: 178-186.

19. Blomberg A (2011) Measuring growth rate in high-throughput growth phenotyping. Curr Opin Biotechnol 22: 94-102.

20. Sparkes A, Aubrey W, Byrne E, Clare A, Khan MN, et al. (2010) Review Towards Robot Scientists for autonomous scientific discovery. Autom Exp 2.

21. Chou HH, Berthet J, Marx CJ (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. PLoS Genetics 5: e1000652.

22. Kiefer P, Buchhaupt M, Christen P, Kaup B, Schrader J, et al. (2009) Metabolite profiling uncovers plasmid-induced cobalt limitation under methylotrophic growth conditions. PLoS One 4: e7831.

23. Anthony C, Zatman L (1964) The microbial oxidation of methanol. Biochem J 92: 614-621.

24. Chan H, Anthony C (1992) The mechanism of inhibition by EDTA and EGTA of methanol oxidation by methylotrophic bacteria. FEMS Microbiol Lett 96: 231-234.

25. Dales SL, Anthony C (1995) The interaction of methanol dehydrogenase and its cytochrome electron acceptor. Biochemical Journal 312: 261.

26. Box GEP, Hunter JS, Hunter WG (2005) Statistics for experimenters: design, innovation, and discovery: Wiley Online Library.

27. Atlas RM (2004) Handbook of microbiological media: CRC.

28. Overmann J (2006) Principles of enrichment, isolation, cultivation, and preservation of Prokaryotes. The Prokaryotes 1: 80–136.

29. Kim P, Kim JH, Oh DK (2003) Improvement in cell yield of Methylobacterium sp. by reducing the inhibition of medium components for poly-β-hydroxybutyrate production. World J Microbiol Biotechnol 19: 357-361.

30. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. Journal of bacteriology 119: 736.

31. Dworkin M, Falkow S (2006) The prokaryotes: a handbook on the biology of bacteria. Springer Verlag. pp. 257-265.

32. Delaney NF, Echenique, J.R., Marx, C.J (2012) Clarity: An Open Source Manager for Laboratory Automation. J Lab Autom: In press.

33. Delaney NF, Kaczmarek ME, Marx CJ (2012) Evaluating sources of biases when estimating microbial growth rates in microtiter plates and development of the open-source program Curve Fitter. PLoS One Submitted.

34. Tsuchiya Y, Nishio N, Nagai S (1980) Medium optimization for a methanol utilizing bacterium based on chemostat theory. Applied microbiology and biotechnology 9: 121-127.

35. Schmidt S (2010) Functional investigation of methanol dehydrogenase-like protein XoxF in Methylobacterium extorquens AM1 [Dissertation]: ETH Zurich.

36. Choi J, Kim J, Daniel M, Lebeault J (1989) Optimization of growth medium and poly-β-hydroxybutyric acid production from methanol in *Methylobacterium organophilum*. Kor J Appl Microbiol Bioeng 17: 392-396.

37. Kensy F, Zimmermann HF, Knabben I, Anderlei T, Trauthwein H, et al. (2005) Oxygen transfer phenomena in 48-well microtiter plates: Determination by optical monitoring of

sulfite oxidation and verification by real-time measurement during microbial growth. Biotechnol Bioeng 89: 698-708.

38. Marx CJ, Miller JA, Chistoserdova L, Lidstrom ME (2004) Multiple formaldehyde oxidation/detoxification pathways in Burkholderia fungorum LB400. J Bacteriol 186: 2173-2178.

39. Morel F, Hering JG (1993) Principles and applications of aquatic chemistry: Wiley-Interscience.

40. Mokhtari-Hosseini ZB, Vasheghani-Farahani E, Heidarzadeh-Vazifekhoran A, Shojaosadati SA, Karimzadeh R, et al. (2009) Statistical media optimization for growth and PHB production from methanol by a methylotrophic bacterium. Bioresource technology 100: 2436-2443.

41. Mokhtari-Hosseini ZB, Vasheghani-Farahani E, Shojaosadati SA, Karimzadeh R, Khosravi-Darani K (2009) Media Selection for Poly (hydroxybutyrate) Production from Methanol by *Methylobacterium Extorquens* DSMZ 1340. Iran J Chem Chem Eng Vol 28.

42. McIntire WS, Wemmer DE, Chistoserdov A, Lidstrom ME (1991) A new cofactor in a prokaryotic enzyme: tryptophan tryptophylquinone as the redox prosthetic group in methylamine dehydrogenase. Science 252: 817-824.

43. Laukel M, Chistoserdova L, Lidstrom ME, Vorholt JA (2003) The tungsten containing formate dehydrogenase from Methylobacterium extorquens AM1: Purification and properties. Eur J Biochem 270: 325-333.

Chapter 3

Evaluating sources of biases when estimating microbial growth rates in microtiter plates and development of the open-source program Curve Fitter.

A description and evaluation of the statistical methods used to ascertain microbial growth rates.

This work was done in collaboration with Maria E. Kaczmarek.

Abstract

A comparison of the growth dynamics of different microorganisms, or the same organism under different conditions, has been fundamental to many research projects. If microtiter plates and laboratory automation equipment are used, one can with relative ease generate thousands of growth curves for microorganisms and look for differences between strains or media conditions. Although finding what has traditionally qualified as a statistically significant differences with such large datasets is easy given their size, it is not clear how best to fit the data that is produced to any particular growth model, or how to account for, and most importantly detect, the systematic biases that might arise and lead to misleading results. In this paper we investigated the practical and statistical issues related to fitting an exponential growth model to optical density data from bacterial cultures growing in 48-well microtiter plates. Using a large database of growth curves of *Methylobacterium extorquens*, we found that over a wide range of optical density values the exponential model was a very good approximation to the underlying behavior of the cultures. Despite this, the inferred growth rate depended upon choices of the inoculum size, blanking strategy, range of data fit and blocking scheme. Furthermore, the error was not normally distributed and so a mixture model was needed to account for outlying data points that can occur. To implement this mixture model and other fitting routines for growth curves, as well as to help assess model adequacy and the discovery of systematic biases we created the program Curve Fitter. Curve Fitter is a visualization tool and scripting environment that allows users to fit multiple models using non-linear routines, perform model comparisons and check model diagnostics such as QQ Plots, residual through time plots and heat maps of different summary statistics across a microtiter plate.

Introduction

In microbiology, a classic way to quantitatively evaluate the performance of a bacterium is to determine how fast it grows in a particular environment. The growth rate of a bacterium is a useful quantity for many reasons. From a physiological perspective, in a planktonic environment, growth rate is the highest level of phenotype that integrates across all underlying phenotypes occurring within the cell. From an evolutionary perspective, the growth rate of a genotype is typically also the major component that determines competitive fitness. Given the central role of growth, it is also commonly used as a metric to assess how much a particular genetic change or environmental perturbation really 'matters' to the cell.

Determining the grow rate of bacteria under some culture conditions is now relatively easy to do in a high-throughput fashion. In a typical growth rate study, different treatments of interest (either different strains or different culture conditions) are compared by obtaining growth curves for each treatment, summarizing each curve by a single estimated growth rate, and then comparing this rate across treatments. Although not required or historically used for this work, multi-well plates, laboratory automation and computer databases can be used to conduct these studies on a large scale. This ability to efficiently collect large amounts of growth data is exciting not only because it makes research more efficient, but also because it allows for new questions to be explored.

However, high-throughput growth rate datasets also present new challenges to how we analyze growth data, and standard statistical tools can be inadequate for the task. For example, one problem is how to determine if two treatments have different growth rates. A sensible approach to this for a study with a limited ability to collect data might be to take a few replicated measurements of growth curves for each treatment, and then to test if the mean growth rate

estimated for each treatment is significantly different by using a standard method of statistical modeling, such as a t-test or ANOVA. This approach however might give a very misleading result when applied to a large dataset of growth rates. The reason is that for typical linear statistical models, such as the t-test or ANOVA, the standard error of estimated differences between treatments constantly decreases with the square root of the sample size. This in turn implies that with a high-throughput system one could detect even the smallest of differences simply by using modern equipment to collect ever larger samples, progressively guaranteeing that any difference would be significant by the typical p-value criteria.

It is usually the case however that when the sample sizes become very large conventional significance criteria like p-values are less relevant. Instead, it is more important to ensure that the effects of small biases, which always exist, are adequately screened for as they might be the only factor explaining an inferred difference between treatments. An example of such bias might be cryptic environmental heterogeneity that affected the growth curves but was unknown, or not properly accounted for, when the experiment was designed. Such biases can later be accounted for if possible causes of bias can be proposed, such as the order in which measurements were taken, and then found to be consequential in a post-hoc analysis that uses graphical or numerical procedures. Having tools to visualize data according to different covariates is important for this purpose, as if sources of bias can be identified and accounted for then both measurements and conclusions can be made more accurately. Sometimes however, no sources of bias can be identified but they be detected by replicating ones experiment and seeing that the new estimates for different effects are outside of the range previously inferred. For example, one study found that replicate runs of a flow-cytometry assay designed to identify mutations with very small effects showed significantly more variation as values were re-estimated than a standard model

would predict [1], and so used hierarchical models to account for these unknown effects and the variation they introduced. In general, the challenge with data from high-throughput measurement systems is not to collect enough data to be able to ensure that a small difference will be statistically significant, but rather to test and verify that the model used to assess significance is, or is not, an accurate approximation to reality.

Another common issue when assessing the differences between treatments using standard tools is that the assumption of a normally distributed error for each observation (or for the batch effects underlying a set of observations) is not robust to outliers, though they are frequently found in high-throughput datasets. As a result, data preprocessing in the form of filtering or excluding observations typically takes place. For example, studies looking at the effects of gene deletions upon high throughput growth data in yeast systematically excluded noisy measurements before proceeding with the analysis [2,3]. Although it is clear that such outlying observations should not be naively incorporated into an analysis, it is less clear how to best identify them to ensure both that the procedure accounting for them has not biased one's conclusions, and that is readily interpretable.

In light of these issues, we investigated methods of estimating microbial growth rates in a high-throughput fashion. In this paper we present two things: a new, open-source software, Curve Fitter, and an analysis of the effect of experimental and analytical choices upon accurate estimations of the growth rate. Studying the growth rate of bacterial cells is a pursuit almost as old as microbiology itself [4]. However, there are a surprisingly large number of ways to estimate parameters from growth curve.

The typical model for growth in batch culture is that when cells from a nutrient limited environment are first placed in fresh media, they begin to undergo a transformation and after a

period of time known as the lag phase, the culture eventually obtains an equilibrium steady-state growth rate defined by an exponential growth equation [4]. Eventually however, the growth of the culture begins to change the environment of the media, and as the cells use up the resources and excrete byproducts, the growth rate slows and the density of the culture stabilizes.

To analyze this growth pattern, there are two general approaches. The first is to attempt to model the entire growth curve, from the lag phase through the stationary phase. Typical mathematical forms to do this include specific functional forms such as the Gompertz [5] , Baranyi [6] or the general Schnute model [7], as well as semi-empirical spline models that interpolate between points [8]. Alternatively, the second approach focuses solely on estimating the steady-state growth rate assuming a constant exponential model. This is most commonly done by taking a logarithmic transform of the data over a fixed range of OD values, regressing on the measurement time, and taking the estimated slope as the inferred exponential rate (such as in [9]). However, more complex variations exist [1]; for example, one study analyzed the log transformed data by taking the slope between every third consecutive measurement, excluding the highest two readings or any above a threshold value, and then taking the mean of the remaining five slopes estimated that way [10].

For this work we focus on estimating the growth rate for the steady state exponential model as the parameter of interest, and the question of determining if two or more strains differ in their estimated rates. Although all models we tested provide very good approximations to the data, we focused on the simple exponential model because it is a parametric form with clear biological interpretability. Our data for this study are a compendium of growth curves generated by a recently described high-throughput automation system [11]**,** which takes periodic OD_{600} measurements (hereafter simply ‘OD’) of growth in 48-well microtiter plates. The organism we

used were strains of *Methylobacterium extorquens* AM1, and aerobic bacterium traditionally studied for its metabolism of single carbon compounds such as methanol and methylamine. The cells were grown in a recently described medium specifically designed to be robust to minor variations in its components so as to minimize media batch effects [12].

We evaluate the issues in estimating the growth rate from two perspectives: the effect of different experimental designs, as well as methods of analysis of the resulting data. If a proper experimental layout is used, we find that the system can very reliably and uninfluenced by bias detect growth rate differences to within 2% using standard linear models after doing growth rate estimation as a preprocessing step. Below this range, we find that the basic model is still useful but more care must be taken in what effects are controlled for and how they are modeled. For analyses trying to detect differences of either large or small effect, we find that occasionally deviant results can greatly skew the analysis, and so conscientious evaluation of the model with diagnostic plots and visual examinations of the data are essential. To implement fitting of the models we consider in this paper, as well as to rapidly allow experimenters to view, fit, and most importantly evaluate model fits with graphical diagnostics, we created the program Curve Fitter, which implements the fitting routines and provides visualizations of the data generated by high throughput growth assays on microtiter plates to allow users to rapidly detect any systematic biases. The program is available for download from www.evolvedmicrobe.com/CurveFitter/index.html, and the website also contains example datasets to show the type of data and analyzes we discuss here.

Results and Methods

Instrumentation and growth conditions – Cultures growing in 48-well plates had their OD measured through time by a recently described robotic system [11] . This consisted of a plate

shaking tower with a de-lidding station (Liconic), a robotic arm (Twister, Caliper), and a plate reader (Victor 2, Perkin Elmer). These hardware were integrated and controlled by the open-source manager program, Clarity [11]. The system is designed to read multiple plates simultaneously by alternating which one is on the plate reader at any moment. A video showing the system and how plates are moved between the reader and the shaker is available at: www.evolvedmicrobe.com/LabAutomation.html. Fig. 3.1 also shows the type of data produced.

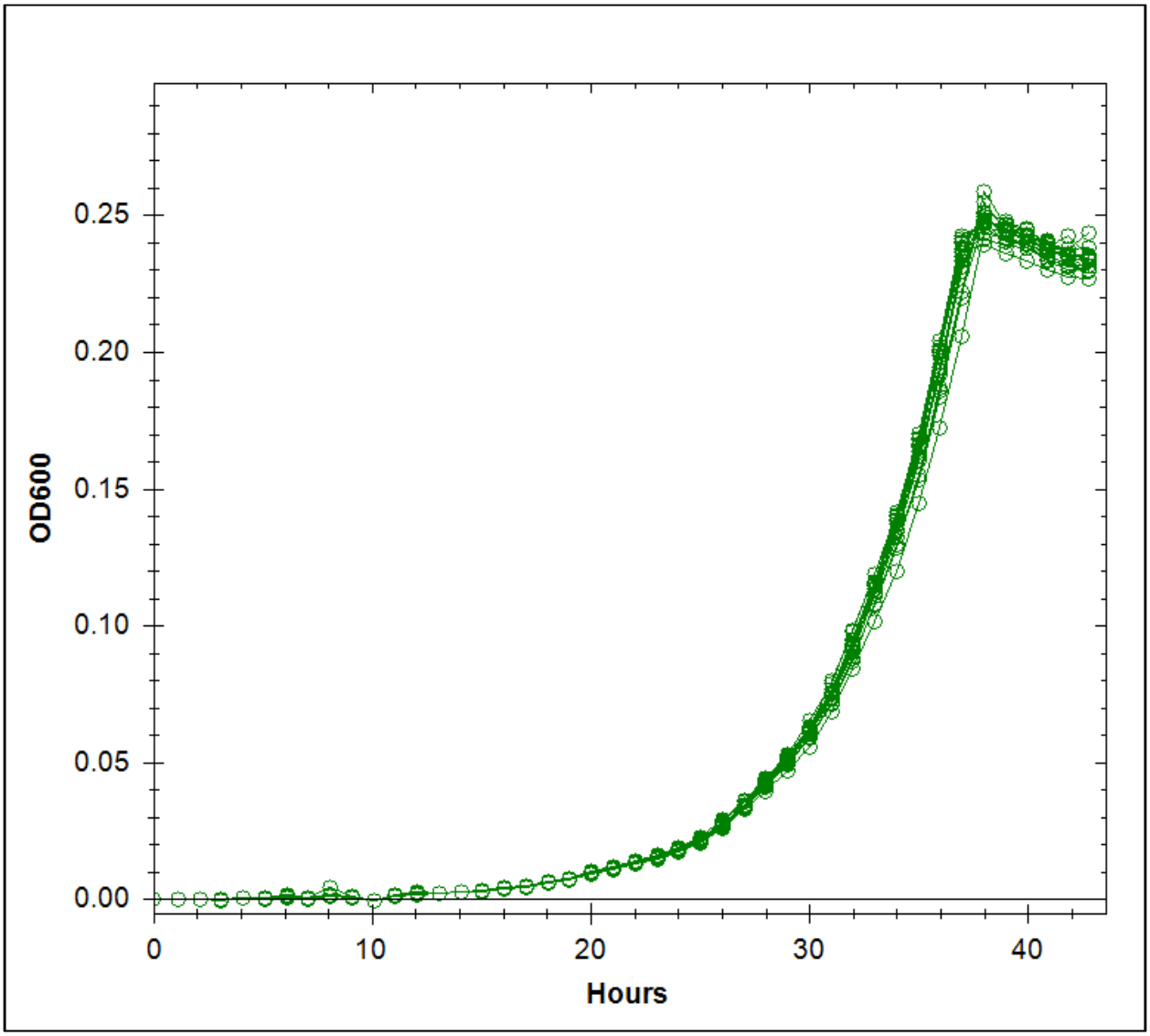


Figure 3.1 – Growth plots of *M. extorquens* growing in 48 well plates.

The *M. extorquens* AM1 cultures were a wild-type strain with a deletion of three genes needed for cellulose biosynthesis ([12]), that was being serially passaged under these growth conditions. Growth was in a in the newly described 'MP' media which was optimized for *M. extorquens* in the accompanying paper ([12]). All cultures reported here were supplied with 20 mM

methylamine HCL as the growth substrate, a concentration which allows for a final OD measurement of ~0.25. Cultures were routinely inoculated from stationary-phase cultures grown under the same conditions (typically in stationary phase for between 6 and 21 hours). The OD was measured at 600 nm every 50 minutes.

The model for estimating growth rates and its statistical properties

Determining a growth model with measurement error – Central to any statistical analysis is correctly modeling the randomization process affecting the observed data, which for growth rate measurements on a plate is largely caused by the noise in the OD measurements. We found that, after removing obvious outliers (by criteria identified in a following section) the errors in the OD readings are well approximated by a normal distribution with a standard deviation of 0.0012. We determined the error model, and also whether the magnitude of the error varied proportionally to the absolute OD reading, by creating a dataset with twice the amount of data required to accurately estimate the growth rate, and used half the data to estimate residual values by predicting points for the second half of the data. We grew strains on either succinate or methylamine (to provide a range of growth rates) in different wells on three different plates. Two readings for each plate were taken 10 minutes apart, with a 20 minute interval before another paired measurement. We then alternatingly divided the readings in to two groups. We used the first set of data points to predict the OD values at a given time using an exponential growth model, and used the second set of data points to determine the residual values at those points. The curves were fit and the residuals were analyzed only when OD readings were between 0.02 and 0.18 after blanking (for reasons discussed later). We further verified that the magnitude of the measurement error did not vary with the OD readings by running replicate

wells filled with varying levels of red wine vinegar which absorbs at OD_{600}, and found that the magnitude of residual error did not scale with the mean reading.

We therefore modeled the OD reading in a well at a point in time according to the following exponential growth equation:

$$OD_i = Ae^{rt_i} + C_i + Error \quad (1)$$

In the model above, r is the growth rate of interest, A is the initial population size, C_i is an offset value and t_i is the measurement time. In this paper, we consider four variations of this model all of which are implemented in Curve Fitter with non-linear fitting routines to find the maximum likelihood parameter estimates. In the first model, henceforth M1, C_i is assumed to be equal to zero for all time points and the error is normally distributed. This can be achieved by "Blanking" the readings as discussed later. The second and third differ in that the C_i is inferred but is either assumed constant (M2) or time varying (M3). The fourth model, M4, is a mixture model for the error discussed further on.

Statistical Issues Related to Inferring the Growth Rate – We derived analytically and through simulations some useful statistical properties of the M1 model. One important conclusion from that work is that higher OD readings, where the signal to noise ratio is the highest, carry the most weight in determining the estimated growth rate. This means that high OD readings are particularly important in the estimation process, and every effort should be made to acquire readings from as high an OD value as is possible without violating the exponential model. Conversely, smaller OD readings relative to the instrument error contain less information and are more susceptible to biases. Another result is that although maximum likelihood estimates from non-linear models can often give biased results, we found that this bias is negligible using

simulations; over a range of parameter settings the mean difference between the true growth rate and the estimated growth rate was only -0.00021, an error of ~0.1%. Additionally, although one could fit the M1 model simply by taking a logarithmic transformation of the OD data and performing a regression, we found that although this method can provide a good fit to the data by such measures as R^2, the results are still less accurate than using a non-linear fitter to fit on the original measurement scale.

Perhaps most importantly, the simulations showed that a downstream analysis of the kind considered previously, which ignores the underlying OD readings and only compares replicate growth curves by using a single growth-rate summary statistic for each curve, is a valid approach. Growth rate estimates derived from data simulated for a particular growth rate were normally distributed around the true value the simulation was based on in accordance with the assumption for linear statistical models.

Analyzing the effect of experimental and analytical choices upon estimations of growth rate

Blanking wells – To accurately estimate the growth rate of a culture in a microtiter plate with the M1 model, the OD measurement of each well must be corrected for the blank measurement (or a constant C_i term in (1)). This is the measurement that would be made if no cells were present in the media and is important because subtracting an incorrect blank systematically biases the results. If too small a value is subtracted it leads to artificially slow growth rates, while if too large a value is subtracted it leads to artificially fast growth rates. This can be seen from (1) by noting that the derivative of the logarithm of the OD instead of being exactly equal to the growth rate r, will equal:

$$\frac{dLog[OD(t)]}{dt} = r\left(\frac{Ae^{rt}}{Ae^{rt}+C}\right) \quad (2)$$

Equation 3 also shows that the influence of an unaccounted for *C* term decreases as the magnitude of the OD readings increase, providing yet reason to favor higher readings. In practice, this *C* term can be accurately estimated by using the first OD reading from a plate, which is appropriate if the initial inoculum of cells is heavily diluted. We demonstrated this by measuring OD on a plate for wells filled with sterile or no media for a 48 hour period, and observed that the OD readings varied systematically by plate position, presumably due to small differences created during the manufacturing process, but each wells value varied only trivially through time with a very small standard deviation of approximately 1.3e-4 (Fig. 3.2). This direct subtraction does not account for the inoculum's contribution to OD, but as the bias scales with the inoculation size, this can be avoided if a large dilution is used so that C is small, which also helps guarantee that the cells have time to acclimate to a steady state value. Our laboratory commonly uses a 1/1000 dilution from a culture with a final OD of approximately 0.25 as a standard assay. If an experiment demands a smaller dilution factor (e.g., 1/64), such that the initial readings are reliably above the noise, than one can simply infer the blank value using the M2 model for a constant offset value in the OD readings.

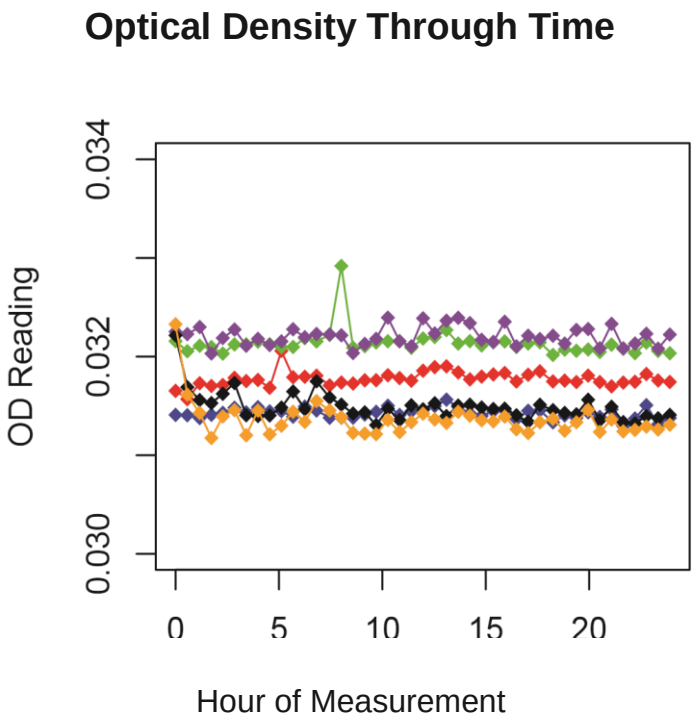


Fig 3.2 - Optical density (OD) measurements over a 24 hour period for 6 randomly selected wells from the same 48 well plate that were filled with sterile media. One well shows the obvious "bump" typical of the type of outlier observation that our mixture model is designed to identify and exclude.

Curve Fitter allows the user to easily blank their data on the first reading or fit the M2 model. We have also observed, both in our system and using other plate readers, that occasionally the very first reading in a plate is discretely different from latter measurements. This is a very clear discrepancy that can occur if the plate is positioned by hand into the system and has not "settled" into a more typical orientation after automated equipment have repositioned it. In the event the first reading is systematically off, we recommend and provide the ability to blank on the second value.

A model for outlier OD readings – We found that the error distribution was only approximately normal after outliers were removed. The Victor2 plate reader has a habit of producing occasional "hiccups" where readings are obviously deviant (as can be seen in Fig. 3.2), but is also prone to the occasional only slightly deviant values whose status as outliers is more debatable. Manually including or excluding such outlier values can be subjective, so we found a way to make this process automatic, quantitatively well founded, and independent of any particular investigative goal by developing a mixture model. We trained this model using a database of growth curves collected as part of ongoing research in the lab. We have been conducting growth curve analyses for the past year and all of these curves are stored in a relational database containing 261 total plates and 564,336 total OD measurements. We queried this database to examine the distribution of all residuals fit with the M1 model without any manual filtering of individual points. The residuals were not generally normally distributed (Fig. 3.3), however the error was well fit by a mixture model of two normal distributions, representing the normal and outlier observations as shown in equation 3.

$$Error = p_1 N(0, \sigma_1^2) + (1 - p_1) N(0, \sigma^2_2) \quad (3)$$

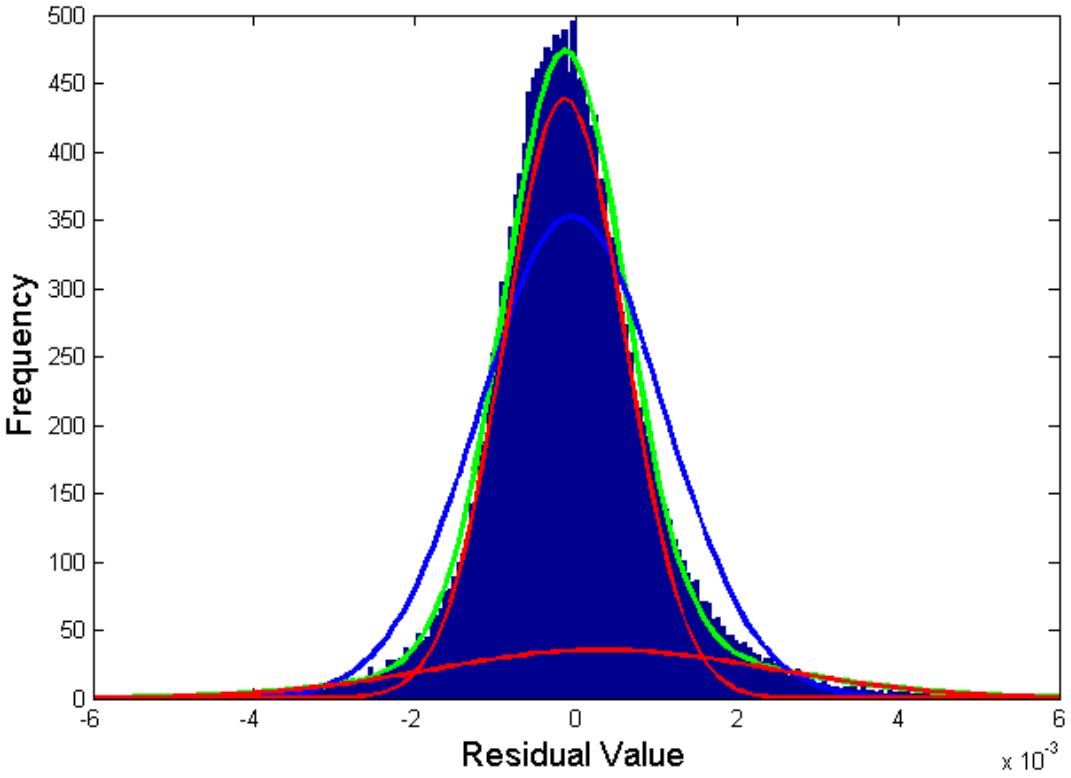


Figure 3.3 – Distribution of residual values after fitting the exponential model. The blue line show the best fit of a standard normal distribution, which is clearly inaccurate. The green line shows the mixture distribution used in fitting, which is composed of the two normal distributions shown in red.

Although other distributional forms like a *t* would likely have yielded similar distributions, we prefer the two-component mixture model form as it provides diagnostics for researchers (Curve Fitter shows the number of outliers in each well position) and allows one to see which OD readings appear to be outliers. Curve Fitter implements this error model and finds the maximum likelihood estimates of the parameters using an expectation maximization algorithm [13] with a non-linear optimization for the M-step. This fitting process is thus an outlier-robust, non-linear regression, where we set the error parameters equal to the estimates from the large database to

avoid the estimation error in refitting them for each curve. Curve Fitter then labels any point with a higher posterior probability of coming from the outlier component in the displayed graphs, and provides summary statistics related to the number of predicted outlier points as part of the display plots as well.

Selecting the range of OD readings to fit and evaluating model adequacy – When fitting the exponential growth models in (1), one must select a range of data to fit, that is a minimum OD value and a maximum OD value over which the exponential model is believed to be valid. As discussed earlier, the higher OD readings are the strongest determinant of the fitted growth rate and have the biggest effect on the accuracy of the measurement. However, if an OD measurement is picked so high that the culture is no longer in exponential growth, this leads to biased estimates, particularly since our equipment only periodically measures each plate, so different plates may have different ODs when measured.

Ideally, almost the entirety of the growth curve over which the OD is noticeably changing should be well approximated by the exponential growth model. Curve Fitter creates plots showing how the estimated growth rate changes through time as different intervals are used to help test these assumptions. One plot shows, for one or many wells, the doubling times obtained at each measured time point by estimating the growth rate using only the two points on either side of that time point (Fig 3.4). This plot is therefore a time series of local growth rate estimates for each data point. If the exponential model is appropriate and a constant growth rate is a good approximation, these interpolated growth rates through time should look like a noisy flat line and not show any systematic bend, which would indicate that the growth rate is slowing down or changing as the OD increases. Additionally, Curve Fitter helps the user evaluate how sensitive any curve is to changing the range of OD values fitted by providing a heat map which shows

how the estimate value for each possible choice of a low and high OD value over which to fit (Fig 3.5). This is provided as a diagnostic and guide in selecting a range to fit. If the exponential model is valid over a given range, this heat map should be relatively stable and not greatly change. For the data shown in Fig. 3.5, the growth rate appears reasonably stable up to an OD of approximately 0.2, and the exponential model appears to be a great approximation.

We emphasize that the exponential model will not always be appropriate for growth curve data, particularly with data from microtiter plates. For example, when we examined data of *M. extorquens* AM1 grown in 96-well plates, it becomes clear that there is no region for which the growth rate is constant; it initially oscillates, and then quickly slows down at higher OD values (leading us to use 48 well plates instead). Although Curve Fitter, like other software, can always fit the model in (1) if it is inaccurate, the graphical diagnostics can alert the user whether the approximation is poor and so this estimated rate is rather meaningless.

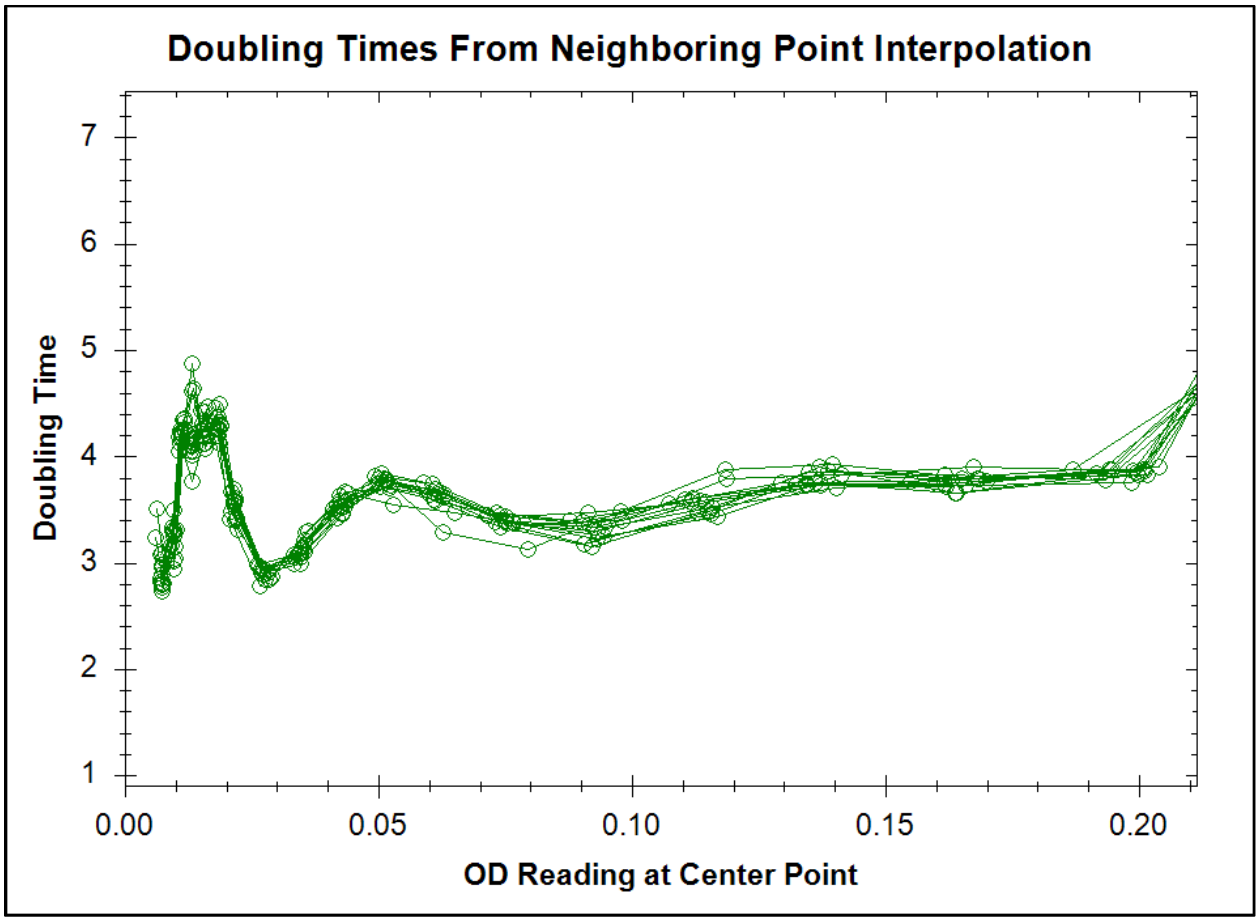


Figure 3.4 – A plot made by Curve Fitter showing the doubling time at each time point estimated using the formula: $\text{Doubling time} = \log(2) / \frac{\log(OD_{t_{i+1}}) - \log(OD_{t_{i-1}})}{t_{i+1} - t_{i-1}}$. The readings eventually stabilize between 0.05 and 0.2, indicating that the exponential model is appropriate. The inset graph shows the actual growth curve this data is derived from.

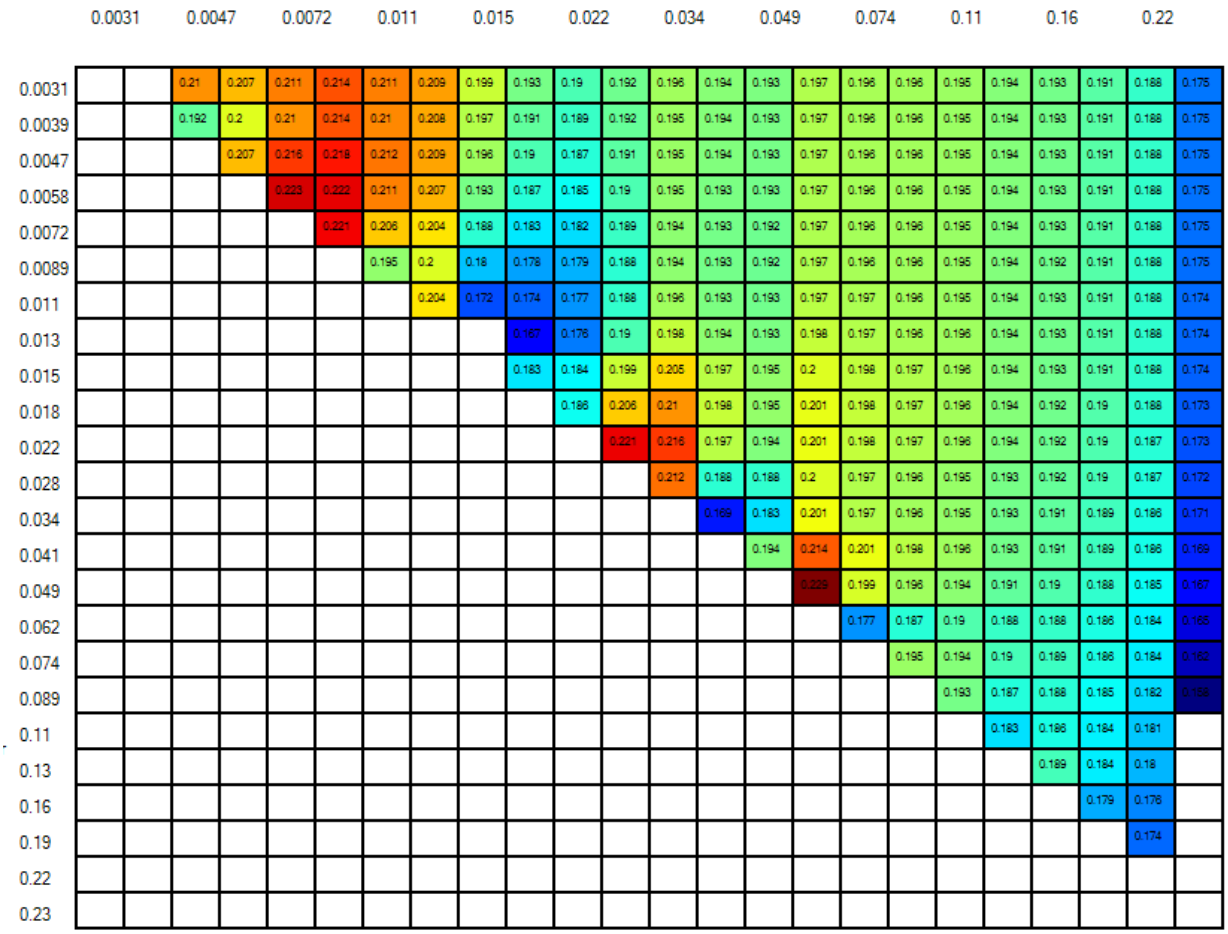


Figure 3.5 – A heat map showing how the estimated growth rate changes as the range of low or high values is fit. The y-axis gives the reading of the lowest value, and the x-axis gives the highest OD value fit. This plot shows the variation in the low range, but the readings become increasingly stable as the OD increases, before significantly slowing down after an OD of 0.2. In accordance with theoretical expectations, the high OD points matter much more than the lower points. Moving left to right along the columns one sees changes in each row, while in contrast moving up down along the rows one sees that the value for any column is more stable.

Using a large database to find and evaluate the effect of model inadequacies – Although the growth curves we collected individually and in small groups appeared to be very well approximated by an exponential model, this model is still of course only an approximation. To look for any systematic deviations from this model and to assess the effect of these on our inferences we looked for deviations from the model with a large database of growth curves that would allow us to descry differences that could not be seen with smaller datasets.

In particular, the salient feature of many alternative models of bacterial growth, such as the Gompertz or logistic models, is that they assume that the growth rate declines as the population size increases. To look for this behavior, we queried, from the same database used to generate the mixture model, any growth curve data where the estimated growth rate was between 0.19 and 0.21, giving a set of 3,045 growth curves evenly centered around a value of 0.2 (this dataset is large because the rate for our most frequently measured strain is ~0.2). From this set, we then estimated for each curve the growth rate at each measured OD by using the slope for the change per time in the log OD values for the 2 points immediately adjacent to that point (as in Fig. 3.4). As expected, the estimated growth rates were more dispersed for the lower readings but then became quite smooth between 0.05 and 0.2, after which they grew noisy again as the cultures left the exponential phase (Fig 3.6). Despite this relative smoothness, examining the mean estimated growth rate at an OD value it was clear that it does noticeably decrease in a linear fashion over the interval from 0.06 to 0.2, and strangely appears to oscillate before this point. A linear regression showed that the mean relative growth rate per hour decreases by -0.01 as the culture increases the OD by 0.1 (Fig 3.6).

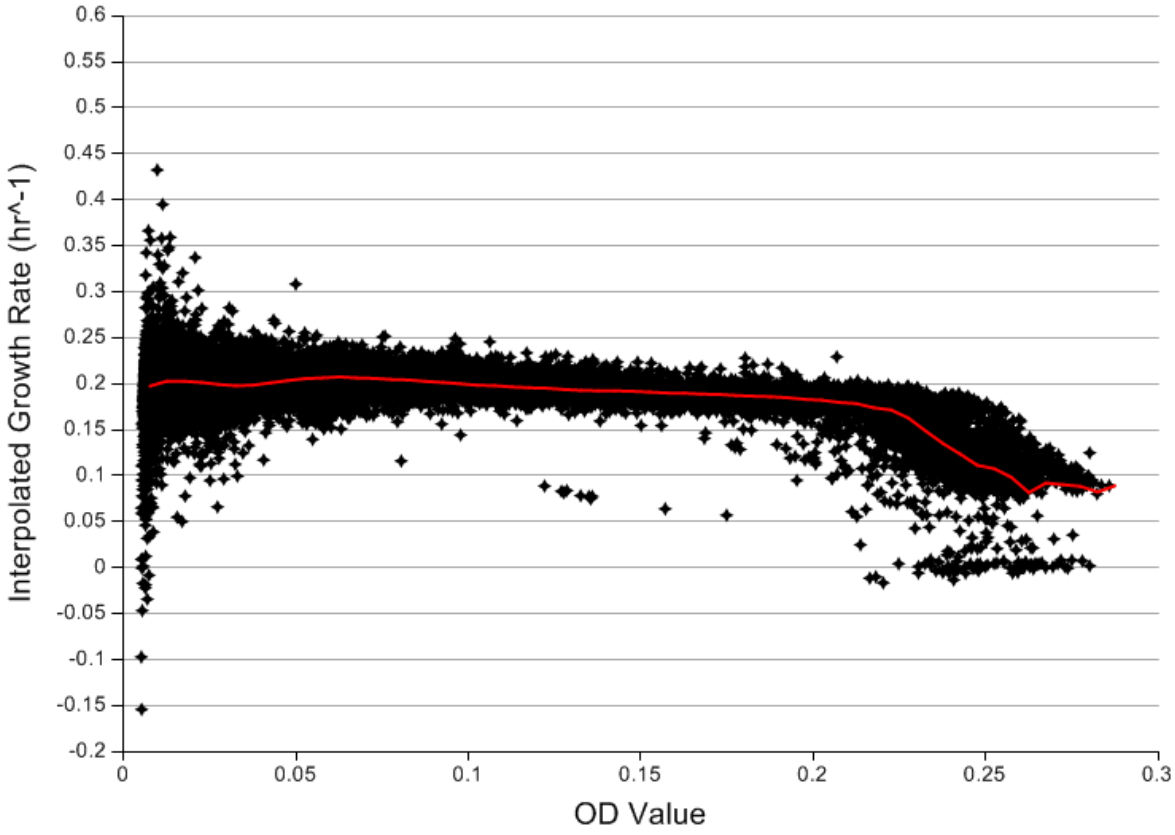


Figure 3.6 – Instantaneous growth rates of 3,045 growth curves of *Methylobacterium* calculated by estimating the rate with the two measurements on either side of a center point. As expected, the wells with higher OD readings showed more stable measurements.

Although this slight decline in instantaneous growth rate that occurs as the OD increases introduces a nearly imperceptible bias into the growth rate estimates, it should be considered if exceptionally small differences are considered important. To model this decline an appropriate model would be one that introduces a new term, *d*, to model the relative growth rate declining as a linear function of the population size:

$$\frac{dOD(t)}{dt} = r_{\max} OD(t)(1 - d \cdot OD(t)) \qquad (4)$$

This equation is simply the logistic growth equation if the decline rate, *d*, is between 0 and 1 (this requirement comes from the fact that the standard presentation of the logistic growth differential equation writes our parameter *d* as the reciprocal of the asymptotic population size, or population carrying capacity, which has to be a positive number). However, we emphasize that if (4) is fit over a narrow range of OD values during growth this parameter is uncoupled from the asymptotic population size and it should not be interpreted as being related to it. In fact, had a standard logistic growth equation been fit to the entire growth curve, the low final OD would have prevented us from obtaining the value of this parameter estimated from the regression using the combined dataset of curves, as it is larger than the reciprocal of what would have been inferred as the carrying capacity.

The model from (4) has fitting routines implemented in Curve Fitter where it is referred to as the Empirical-Logistic to emphasize its separation from the standard interpretation. However, we recommend against the model in (4) as it introduces an unnecessary difficulty for all but the smallest of differences between treatments. If the growth rate of a culture is not constant through time, as in (4), how does one interpret a comparison of two or more strains based on this model? What if one strain appears to have a higher maximum rate, but decreases its growth rate more rapidly than another strain with a slower, but more consistent rate. How can one summarize the physiological effect of a mutation on the growth rate? Instead of the clear parametric picture in (1) where we estimate how fast a strain grows, (4) poses a more complicated question.

Fortunately, empirically there is no added value to using (4) over the basic approximating models in (1). Although fitting the model in (4) can often decrease the observed residual values relative to those obtained by fitting the standard exponential model, based on the curves in our database, if both models are fit they give the same answer. Although the growth rate is constantly

changing in (4), a sensible comparison of the results from the two models can be made by comparing the growth rate derived from (4) at the midpoint of the OD range fit, and the constant growth rate estimated by the exponential model. Comparing these quantities across this dataset, we found that the two are very well linearly correlated (R^2=0.77). Additionally, the presumed advantage of using (4) would be to prevent a bias from being introduced if one treatment of cells happened to have its last OD measurement before stationary phase taken at a lower OD than another treatment of cells, and would thus appear to have a higher growth rate. This is undoubtedly true, if one takes the database curves with nearly equivalent growth rates and models the inferred growth rate as a linear function of the highest OD region included in the range of data to be fit, a significant term for a decrease appears with a p-value below 1×10^{-16}. However, the R^2 for the regression is only 0.068, meaning a trivial portion of the variance has been explained.

An explanation for why the exponential model seems sufficient, as well as an estimate of the level at which it becomes important to consider the richer model in (4), can be derived relatively easily. The bias introduced by approximating (4) with the simple exponential model is created when curves have their last measurement taken at different OD values, so that estimates from some curves include more of a slower period of growth than others. However, with frequent OD measurements all treatments are effectively measured at the same values which negates this effect. It is also easy to calculate the longest expected difference possible for a particular set of experimental parameters, which is the largest OD difference possible for curves sampled at different times. For example, with our dataset, if all the data with an OD above 0.2 are excluded from fitting, if readings are taken every 50 minutes and if the cells double approximately every 3.5 hours, then using the estimated value of *d* for model (4) and assuming no measurement error,

the largest possible bias between two curves should be around 1.5%. This reinforces the fact that if differences below 2% are important, one should both decrease the time between OD measurements and consider richer models, but above that the approximating exponential model is perfectly adequate. Additionally, since we recommend comparing only treatments measured on the same plate (and thus at the same times), in practice the bias is much less.

Although, we do not consider that the richer model in (4) will typically be useful for presenting results, comparing a richer model to a simpler model is a classic method to detect problems with the data or the model. For this reason, Curve Fitter not only allows comparisons to be made between the standard exponential model and (4), but also provides fitting routines for a simple linear three parameter model (5) shown below and referred to as the quadratic model.

$$Log(OD(t_i)) = A + Bt_i + Ct_i^2 \quad (5)$$

The console application in Curve Fitter makes it easy to fit all three of these models and verify that any results are either insensitive to model choice or to examine discrepancies between them.

Blocking structure to avoid plate-to-plate or day-to-day biases – Another experimental issue when performing growth curve analyses in microtiter plates is how one blocks for the effects of the different well positions and different plates. We found that the different wells do significantly differ, but only very slightly, and that replicating across plates was a more important covariate. Fig 3.7 shows the mean growth rates for our most frequently measured strain at every well position. The corner wells are known to have more evaporation and appeared to have slightly slower rates so we typically do not place replicates in them. Growth in the last column also appears to have noticeably slower growth, which we currently attribute to the nature of the airflow around the shaking incubator as we have not observed continued to observe this effect

since placing a cover on the backside of the shaking incubator tower. A more drastic difference was seen between replicate measurements on different plates, but typically even these differences were less than 1%. We believe these differences are caused by small variations in measuring times, as well as possible temperature differences at different positions in the incubator. To control for this bias, we always include the plate as a covariate in any downstream linear modeling. We also try to replicate any treatment to be tested at least 3 times on 3 different plates in different well positions.

Figure 3.7 – Mean growth rates per hour of replicated cultures of our most frequently measured bacterial strain in different wells on a 48 well plate. Systematic variation exists, with the corner wells and the last column showing slower rates, while the center wells are more stable. The middle wells are more equivalent.

Discussion

The exponential growth model is like models of bacterial growth with more parameters in that, at some level of precision, it is wrong. However, unlike richer models, it summarizes growth with a single interpretable parameter that communicates how fast a bacterium grows in a particular environment. We evaluated the ways in which this simple model failed to capture aspects of the growth process in microtiter plates or was systematically biased. We found that if an experiment uses a large dilution of cells, replicates across plates and produces data that passes graphical diagnostics, then the model is an incredibly good approximation to reality. It provides robust and meaningful results which are almost certainly correct if the estimated effect size is over 2%. Below this 2% threshold more care should be taken, and if small differences of less than half a percent are important, a richer model should be used regardless of the sample size obtained. These cut off values are of course approximate, and for any case should be interpreted in light of the issues discussed here.

With large datasets, differences that are difficult to detect with smaller datasets can become noticeable. A strange behavior we found with our large collection of growth curves is that at lower OD readings, the value of the relative growth rate appears to undulate (Fig. 3.4). This pattern is very hard to discern from any single curve as these low OD values are also the readings with the highest noise, but it is clearly evident after examining hundreds of similar measurements. These early undulations do eventually settle into a nice monotonic function (Fig 3.4), and so we fit only over this consistent range to avoid the variance introduced during this initial stage of growth. One possible explanation for the behavior is that these post-transfer undulations might reflect synchronous cell divisions that initially occur as the culture leaves lag phase but that eventually become asynchronous overtime. Another possibility is that the optical

properties of the cells are changing and have not reached a steady-state during the first several divisions. Although this entire paper was about the growth of cells, it should be remembered that all of the data are actually measurements of the light scattering properties of liquid volumes in a microtiter plate. It is possible that for some types of bacteria or stages of growth this could be an important distinction.

In addition to allowing one to detect subtle and systematic effects, from a practical standpoint, large datasets also make it difficult to carefully evaluate and monitor the quality of the data collected. Like many labs, our system uses microtiter plates to collect a large amount of data; it is able to take measurements from over 1,500 wells at any one time. Curve Fitter was specifically designed to aid with the organization, visualization and quality control of this data. Although many programs can fit growth curves and are designed with non-linear fitting routines, few are specifically designed to aid in the visualization and quality control of data produced by microtiter plates. Curve Fitter allows users to quickly and easily assess every one of the assumptions and issues discussed in this paper and also implements the mixture model to account for the error process. Quantile-Quantile (QQ) plots and residual through time plots ensure data quality, model adequacy and help identify deviant behavior, heat maps of the summary statistics for different positions in the microtiter plates are available, and plots of the fits compared to the data for each well are readily scrolled through. The program also includes graphical visualizations of sensitivity analyses of the range of data fit over. Examining these diagnostic plots over the years, we have been able to detect a great number of things that matter more than any statistical nuance. We have seen the effects of one instrument being slightly perturbed so that the bottom of only row of a microtiter plates is scratched on every OD reading, as well as signs that from when the

humidifier in the room begin to periodically fail, or know when the light bulb in the plate reader needs to be changed.

Although Curve Fitter is composed of a graphical user interface to help with many common tasks, perhaps more importantly, the classes, namespaces and methods of Curve Fitter are directly accessible from a numeric scripting environment that is incorporated into the program as a console application. The most appropriate analysis or question for a dataset of growth curves may not always be the same, and so Curve Fitter was designed to be easily extendable by anyone familiar with basic computer programming. Curve Fitter includes a command shell that is built on top of the Sho Playground for Data. This is a Python language interpretive console that includes many of the numeric and plotting features of a program like Matlab. Within it, users can quickly and with a powerful object-oriented language call all of the classes in Curve Fitter and Sho, to import or manage data, plot it, show histograms, generate visualizations of patterns across a microtiter plate or compare values to any number of simulated or calculated statistical distributions easily. This makes it simple to quickly fit all of the models considered in this paper, as well as others not discussed, and ensure that the results are robust to model choice. It also makes it easy to determine if a model is not a good approximation for the observed growth dynamics.

Analyzing microbial growth curves at one level is simply asking how the biomass of a culture changes through time. However, there are a great many ways to ask and answer this question, and no one method is appropriate for every different objective. In formulating any statistical model to evaluate ones conclusions, there is often a bias versus variance tradeoff. A model that captures all aspects of reality may have too many parameters for any sensible inferences to be made, while simple models might give biased results because they have left out important factors

that affect the data. Similarly there is often a rigorousness versus interpretability trade-off as well. Although exceptionally complex models or methods of inference that account for every aspect of the data collection process may be the most rigorous, they can also be unnecessarily complex and obfuscate the meaning of the data that was collected.

After evaluating many growth curves in microtiter plates, we found that for the growth of *M. extorquens* in 48-well plates, the decision point between more complex models and the simple exponential model is an effect size of roughly 0.5-2%. Above this point comparisons with linear models are perfectly adequate, though one should include both an effect for each microtiter plate and the day the curve was measured on as blocking covariates. This approach might not be appropriate for all growth curves datasets created by using microtiter plates. However, we believe the visualization and analysis tools available in Curve Fitter will be. Science is an endless cycle of identifying problems in models and then adjusting the models to account for them, and being able to plot and calculate how data differs from expectations has always been an important part of this process.

Acknowledgements

We would like to thank Martin Lysy for helping to track down a bug in the code implementing the EM algorithm in Curve Fitter. Lon Chubiz and members of the Marx lab provided excellent feedback on an earlier version of this manuscript. N. F. D. was supported by an NSF graduate student fellowship and this research was funded by an NSF CAREER grant (DEB-0845893).

References

1. Blomberg A (2011) Measuring growth rate in high-throughput growth phenotyping. Current Opinion in Biotechnology 22: 94-102.

2. Agrawal AF, Whitlock MC (2011) Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics 187: 553-566.

3. Manna F, Gallet R, Martin G, Lenormand T (2012) The high-throughput yeast deletion fitness data and the theories of dominance. Journal of Evolutionary Biology.

4. Neidhardt FC (1999) Bacterial Growth: Constant Obsession with dN/dt. Journal of bacteriology 181: 7405-7408.

5. Li G (2011) Optimal and efficient designs for Gompertz regression models. Annals of the Institute of Statistical Mathematics: 1-13.

6. Grijspeerdt K, Vanrolleghem P (1999) Estimating the parameters of the Baranyi model for bacterial growth. Food microbiology 16: 593-605.

7. Schnute J (1981) A versatile growth model with statistically stable parameters. Canadian Journal of Fisheries and Aquatic Sciences 38: 1128-1140.

8. King RD, Rowland J, Oliver SG, Young M, Aubrey W, et al. (2009) The automation of science. Science 324: 85-89.

9. Bollenbach T, Quan S, Chait R, Kishony R (2009) Nonoptimal microbial response to antibiotics underlies suppressive drug interactions. Cell 139: 707-718.

10. Warringer J, Blomberg A (2003) Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae. Yeast 20: 53-67.

11. Delaney NF, Echenique, J.R., Marx, C.J (2012) Clarity: An Open Source Manager for Laboratory Automation. Journal of Laboratory Automation: In press.

12. Delaney NF, Kaczmarek ME, Ward LM, Swanson PK, Lee MC, et al. (2012) Development of an optimized medium, strain and high-throughput culturing methods for *Methylobacterium extorquens*. PLoS One Submitted.

13. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological): 1-38.

Chapter 4

The distribution of beneficial fitness effects for a complex quantitative trait, the growth rate of a bacterium, is skewed towards large effect mutations that occur at high rates.

An empirical measurement of the distribution of fitness effects using the tools developed.

This work was done in collaboration with Maria E. Kaczmarek and Paige K. Swanson.

The most basic, and possibly most important, things one needs to know in order to predict and understand how a population of organisms will adapt, is what beneficial mutations are available to those organisms and what their chances of occurring are. This distribution, typically called the distribution of beneficial fitness effects, or the DBFE, is centrally important to much of evolutionary theory [1]. The DBFE and the associated rate at which mutations appear determines how the rate of adaptation of a population size changes as it varies [2], whether recombination can speed adaptation [3,4], whether adaptation is likely to be due to many or a few genes [5,6] and whether a population can avoid extinction in a changing environment by constantly adapting [7].

Despite their central importance, there is little general agreement about the rate of beneficial mutations (U_b) or the nature of their distribution. There are two main reasons for this. First, although there are theoretical reasons to expect that general characteristics of the DBFE might be invariant to the biological specifics of an evolving system [1,8,9,10,11], others have argued that these quantities are intimately associated with idiosyncrasies of the genotype and physiology of the organism being considered [6,12]. The typical pattern of adaptation may also depend on factors such as the relative fitness of the current genotype of the organism [11,13], or the complexity of the genetic architecture underlying the trait being selected for [6].

The second reason is due to the experimental difficulty in measuring the DBFE, even after conditioning on a particular genotype evolving in a particular environment. There are relatively few studies that have attempted to completely characterize DBFEs and this has impeded efforts to find or verify a general pattern for the DBFEs that might be relevant to any particular situation. There are multiple reasons why characterizing the DBFE is difficult. First, beneficial

mutation rates are typically very low, meaning that one must have a mechanism to screen thousands to billions of mutations in order to find a single beneficial one, a task typically impossible for organisms with long generation times. A few studies using bacteria have attempted to circumvent this difficulty by linking fitness to a genetic marker that can be easily screened for using antibiotics. This allows an entire population of cells to be quickly reduced down to only those carrying a beneficial mutation. For example, McDonald et. al. investigated the fitness effects of beneficial mutations that increased expression tied to a specific promoter by placing that promoter in front of a kanamycin-resistance gene [12]. Other studies have simply made resistance to the antibiotic the trait that is being selected for. This approach was taken by Schenk et. al. who used error-prone PCR to generate mutations in an enzyme that could induce resistances to a novel antibiotic. The great strength of these studies is that they allow for very accurate sampling of the DBFE. However their design suffers from the drawback that any beneficial mutations must be constrained *a priori* to a specified trait that can be linked to a binary marker system. One must then also be able to map this discrete trait, such as the ability to grow on antibiotic at a particular concentration, to some continuous and meaningful measure of fitness.

The most common method for finding beneficial mutations in microorganisms avoids this complication by using an evolution experiment to screen for beneficial mutations. This is typically done by seeding a population with isogenic strains, propagating it for some time, and then selecting adaptive mutations that although initially rare, have risen to high frequency by virtue of their selective advantage. Such experimental evolution studies have provided enormous insights into evolutionary processes [14] and the ability to sequence the entire genome of the strains isolated from these experiments has also provided examples of mutational effect sizes,

rates of introduction and epistatic interactions between beneficial mutations in these populations [14,15,16,17,18].

However, the beneficial mutations obtained from typical experimental evolution studies provide surprisingly little information about the actual DBF. Most of these experiments have been conducted by evolving populations that have so many individuals in them, that in the course of a few generations not just one, but multiple beneficial mutations will appear and escape stochastic loss (i.e., drift) in the population [4,19]. This means that the mutations that eventually rise to a high enough frequency to be detected or sampled must both escape drift and outcompete the other beneficial mutations that have occurred and are simultaneously increasing in frequency. As a result, these "winner" mutations are drawn from the tail end, or the extreme, of the DBFE, and it is a well-established principle that the distribution of extreme values is largely independent of the underlying distribution of all possible values [20]. The independence of the extremes from the bulk of the distribution introduces a simplification whereby their dynamics in large populations can be effectively modeled by only considering a single selective coefficient that is representative of the tail values of the distribution that commonly occur at that large population size [2,21,22]. That is, although in large populations one cannot know the complete DBFE, for many purposes this also means one does not need to, as only the tail of the distribution is relevant.

That the mutations recovered in evolution experiments using large populations only represent the extremes of the DBFE can be problematic for reasons unrelated to the difficulties this creates in characterizing the DBFE. Experimental evolution is not only used to investigate fundamental principles of adaptation, but is also increasingly used to find beneficial mutations that allow a

bacterium to better serve an applied purpose or to provide insights into the genetics underpinning a particular physiological trait [23,24]. For example, two recent studies [25,26] evolved *E. coli* populations in order to select for tolerance to isobutanol, a promising biofuel whose current production is limited because it is toxic to the bacteria that can produce it. When large populations are evolved for such projects, one could suspect that the beneficial mutations at the tail end of the DBFE that are recovered in these experiments, or the "hopeful-monsters" in the earlier parlance of the field, are very atypical compared to those that might appear in more common adaptive scenarios. This may give them undesirable characteristics.

In particular, mutations that induce large-scale and drastic global regulatory changes are frequently the earliest to arise in evolution experiments with bacteria [17]. This has been true for experiments evolving *E. coli*, where, because the majority of evolution experiments have worked with large populations of this organism, this general pattern can be seen. Regularly in these experiments, mutations have been found that affect such large-scale regulators as *spoT*, *rpoS* and even the core RNA polymerase genes *rpoB* and *rpoC* [17,27,28,29,30]. Mutations to these central nodes of cell physiology effect many things simultaneously. Conrad et. al. [27] investigated such effects by reintroducing several mutations in *rpoB* and *rpoC* that arose during an evolution experiment into the ancestral genome. They found that in addition to providing a very large 60% increase in growth rate, on average each one affected the transcription level of over a thousand different genes. This ability to reprogram so much of cellular physiology with only one mutational step may likely be a criteria for mutations that are found on the tail end of the DBFE that is recovered in these experiments.

However, although *rpoB* is mutated in laboratory experiments of *E. coli* where it can be beneficial, a recent study that examined the relative ratios of non-synonymous to synonymous nucleotide diversity for hundreds of genes and many species of gut bacteria found that *rpoB* was the single most conserved gene in the human gut microbiome. One might expect that loss-of-function mutations to such unchanging facets of the cellular architecture could exert strong pleitropic effects, and these might influence the ultimate adaptive trajectory of the population being evolved [31,32]. Herring et. al. [28] demonstrated this possibility when they found that mutations that arose in a selection experiment for growth on glycerol were only beneficial in the presence of earlier mutations that had already appeared on the genetic background, and were thus simply compensating for pleiotropic side-effects.

The general prevalence of such epistatic interactions can be subject to debate, but two aspects of evolutionary adaptation, particularly for quantitative traits, appear to be accepted. The first is that the density of the DBFE should decrease with the effect size of the mutation, so that small effects are more prevalent than large. The second is that evolution in small populations should lead to evolutionary outcomes that are more variable when replicated and involve beneficial mutations with smaller effects than large populations. Rare, but large-effect mutations, are more likely to appear in large populations and once they do they can largely determine the dynamics of adaption.

In this paper, we tested both of these assumptions. Whether mutations with smaller effects on fitness were more common than those with larger effects, and whether evolution experiments conducted at smaller population sizes would show a greater diversity of outcomes than those conducted at larger population sizes. To do this, we evolved populations of a bacterium,

Methylobacterium extorquens AM1, at two different population sizes. Our expectation was that evolved isolates obtained after evolution at large population sizes should have a higher mean fitness and be physiologically more similar than those obtained after evolution at small population sizes. We found this expectation was quantitatively correct, the large populations evolved faster and did have higher mean fitness at the end of the experiment. However, it was qualitatively wrong. Beneficial mutations occurred at a much higher rate and had much higher effects than expected, leading to convergence in both fitness gains and the correlated traits affected by selection across the two population size treatments. Collectively, these data allowed us to for the first time infer the complete DBFE, revealing that there appear to be more large effect beneficial mutations available to the ancestor than those of moderate effect.

Methods and Results

Evolution Experiment

We evolved experimental populations of the aerobic bacterium *Methylobacterium extorquens* AM1, a model organism for the study of C_1 metabolism [33], at two different population sizes. The large population size treatment had an effective population size, $N_e = 1.64 \times 10^7$ while for small population size treatment $N_e = 8.6 \times 10^4$. Population sizes are effective for the probability a beneficial mutation escapes stochastic loss [34]. For comparison, the classic evolution experiments of Richard Lenski used a regime where $N_e = 3.3 \times 10^7$ [35], roughly double our large population size.

Full details of the evolution experiment are provided in the SI. Briefly, in order to sample a large number of mutations from the DBFE, we evolved a total of 192 populations, evenly divided

amongst the large and small population size treatments. All populations were grown in 48-well microtiter plates in a shaking incubator at 30 °C, conditions that ensure adequate mixing and stable growth conditions [36]. As small-effect beneficial mutations can take a very large number of generations to reach appreciable frequencies ($\propto \frac{1}{s}$) and because cross-contamination can occur between wells during evolution experiments in microtiter plates [37], we founded all populations using barcoded strains that allowed us to detect any contamination that could occur over the course of a long evolution experiment. Half the wells in each 48-well plate containing the evolving populations were also left empty in a "checkerboard" pattern to minimize contamination. Each population was founded with a 50:50 ratio of strains that each expressed one of two fluorescent proteins, either mCherry or Venus, and also contained one of several uniquely synthesized 44 bp sequences that we inserted in to their genomes so that strains could later be specifically identified using DNA sequencing (Table 4.1, 4.2). In order to enhance reproducibility of the fitness assays and ensure that the environment was stable throughout the experiment, populations were grown in a recently described minimal medium designed to provide adequate nutrients and stable growth even if its component ingredients were largely varied as it was remade [36].

Table 4.1 - Barcoding sequences that were synthesized and inserted into the *M. extorquens* AM1 genome**.** These were used to identify individual strains, check for contamination in the experiment and remove the *cel* operon, which allows for more reproducible growth as described in chapter 3. Each sequence follows the same pattern. There are two flanking regions common to all barcodes (highlighted in blue) that match portions of the *M. extorquens* AM1 genome and are used to identify the recombination site for insertion of the barcode (and simultaneous removal of the *cel* operon). Between the consensus regions is a uniquely synthesized sequence specific to each barcode. All sequences with the exception of a "null" variant contain an *Eco*RI recognition site (highlighted in yellow) so that the type of barcode present in a strain can be identified by either sequencing or a restriction digest. The *Eco*RI recognition sites are spaced 2 bp apart and the relative position of the site was used to name each barcode type. The sequences for the basepairs in the uniquely synthesized region that are not part of the EcoRI site were selected to introduce multiple stop codons. The sequence highlighted in blue on the left side of the barcode represents the AM1 genome up to position 1,217,840 and on the right side it represents the reference genome starting at position 1,225,032. The E-Null barcode type is the original sequence used to first construct the *cel* operon deletion in chapter 3, and so does not have a unique barcode

Table 4.1 (Cont.)

Barcode Name	Sequence
E-Null	GTGAACGCATCCGGAAATCGAGATC----------TCATATGCACGAAGTGAGGACGTTCGATCTCTACA
E-0	GTGAACGCATCCGGAAATCGAGATCAACCGAATTCTCATTAATCAGCTAGTGACTACTCAGCTAGTGGTGAAGTGAGGACGTTCGATCTCTACA
E-2	GTGAACGCATCCGGAAATCGAGATCAAGGTCGAATTCATTAATCAGCTAGTGACTACTCAGCTAGTCCTGAAGTGAGGACGTTCGATCTCTACA
E-4	GTGAACGCATCCGGAAATCGAGATCTTCCTCATGAATTCTAATCAGCTAGTGACTACTCAGCTAGAGGAGAAGTGAGGACGTTCGATCTCTACA
E-6	GTGAACGCATCCGGAAATCGAGATCTTGGTCATTAGAATTCATCAGCTAGTGACTACTCAGCTAGACCAGAAGTGAGGACGTTCGATCTCTACA
E-8	GTGAACGCATCCGGAAATCGAGATCAATTTCATTAATGAATTCCAGCTAGTGACTACTCAGCTAGCTTCGAAGTGAGGACGTTCGATCTCTACA
E-10	GTGAACGCATCCGGAAATCGAGATCCCTTTCATTAATCAGAATTCGCTAGTGACTACTCAGCTAGCAACGAAGTGAGGACGTTCGATCTCTACA
E-12	GTGAACGCATCCGGAAATCGAGATCGGAATCATTAATCAGCGAATTCTAGTGACTACTCAGCTAGACCAGAAGTGAGGACGTTCGATCTCTACA
E-14	GTGAACGCATCCGGAAATCGAGATCCCAATCATTAATCAGCTAGAATTCGTGACTACTCAGCTAGGTTGGAAGTGAGGACGTTCGATCTCTACA
E-16	GTGAACGCATCCGGAAATCGAGATCACCATCATTAATCAGCTAGTGAATTCGACTACTCAGCTAGGGTTGAAGTGAGGACGTTCGATCTCTACA
E-18	GTGAACGCATCCGGAAATCGAGATCAGGATCATTAATCAGCTAGTGAGAATTCCTACTCAGCTAGCCTTGAAGTGAGGACGTTCGATCTCTACA
E-20	GTGAACGCATCCGGAAATCGAGATCTCCTTCATTAATCAGCTAGTGACTGAATTCACTCAGCTAGGGAAGAAGTGAGGACGTTCGATCTCTACA
E-22	GTGAACGCATCCGGAAATCGAGATCCTTCTCATTAATCAGCTAGTGACTACGAATTCTCAGCTAGCCAAGAAGTGAGGACGTTCGATCTCTACA
E-24	GTGAACGCATCCGGAAATCGAGATCGAAGTCATTAATCAGCTAGTGACTACTCGAATTCAGCTAGAATTGAAGTGAGGACGTTCGATCTCTACA
E-26	GTGAACGCATCCGGAAATCGAGATCGTTGTCATTAATCAGCTAGTGACTACTCAGGAATTCCTAGAAGGGAAGTGAGGACGTTCGATCTCTACA
E-28	GTGAACGCATCCGGAAATCGAGATCTGGTTCATTAATCAGCTAGTGACTACTCAGCTGAATTCAGTTCCGAAGTGAGGACGTTCGATCTCTACA
E-30	GTGAACGCATCCGGAAATCGAGATCCAACTCATTAATCAGCTAGTGACTACTCAGCTAGGAATTCTTGGGAAGTGAGGACGTTCGATCTCTACA

Table 4.2 -List of 34 strains created for this evolution experiment. Each row gives the strain ID as well as the barcode sequence and type of fluorescent protein expressed by the strain. All Cherry strains were derived from CM1175, and all Venus strains came from CM1179.

Fluorescent Protein	Barcode Type	Strain Identifier
Cherry	E-Null	CM3120
Cherry	E-0	CM3121
Cherry	E-2	CM3122
Cherry	E-4	CM3123
Cherry	E-6	CM3124
Cherry	E-8	CM3125
Cherry	E-10	CM3126
Cherry	E-12	CM3127
Cherry	E-14	CM3128
Cherry	E-16	CM3129
Cherry	E-18	CM3130
Cherry	E-20	CM3131
Cherry	E-22	CM3132
Cherry	E-24	CM3133
Cherry	E-26	CM3134
Cherry	E-28	CM3135
Cherry	E-30	CM3136
Venus	E-Null	CM3140
Venus	E-0	CM3141
Venus	E-2	CM3142
Venus	E-4	CM3143
Venus	E-6	CM3144
Venus	E-8	CM3145
Venus	E-10	CM3146
Venus	E-12	CM3147
Venus	E-14	CM3148
Venus	E-16	CM3149
Venus	E-18	CM3150
Venus	E-20	CM3151
Venus	E-22	CM3152
Venus	E-24	CM3153
Venus	E-26	CM3154
Venus	E-28	CM3155
Venus	E-30	CM3156

The large population was evolved for 126 generations and the small population evolved for 252, after which all samples from each population were plated to obtain single isolates from each population. For fitness contests, the inference of the DBFE, and later assays, only a single isolate was used from each population.

Estimating the DBFE and beneficial mutation rate

We used the fitness data obtained from the isolates randomly sampled from each well at the end of the experiment to infer the DBFE and the genomic beneficial mutation rate (U_b). Rather than fit any particular parametric form of a DBFE to the data, we fit the data to a discrete distribution that could approximate any continuous distribution (Fig. 4.1). In this model, all beneficial mutations are drawn from a finite number of fitness classes, denoted c_i, $i = 1 \ldots N$. The mutations in each class have the same exact fitness effect and these effects are equally spaced along the number line. As this model specifies a fixed number of fitness classes and fitness values, the only parameters left to infer for it are the total genomic beneficial mutation rate, U_b, and the vector of probabilities for each type of beneficial mutation class, $\boldsymbol{p} = \{p_1 \ldots p_N\}$, where $1 = \sum_{i=1}^{N} p_i$.

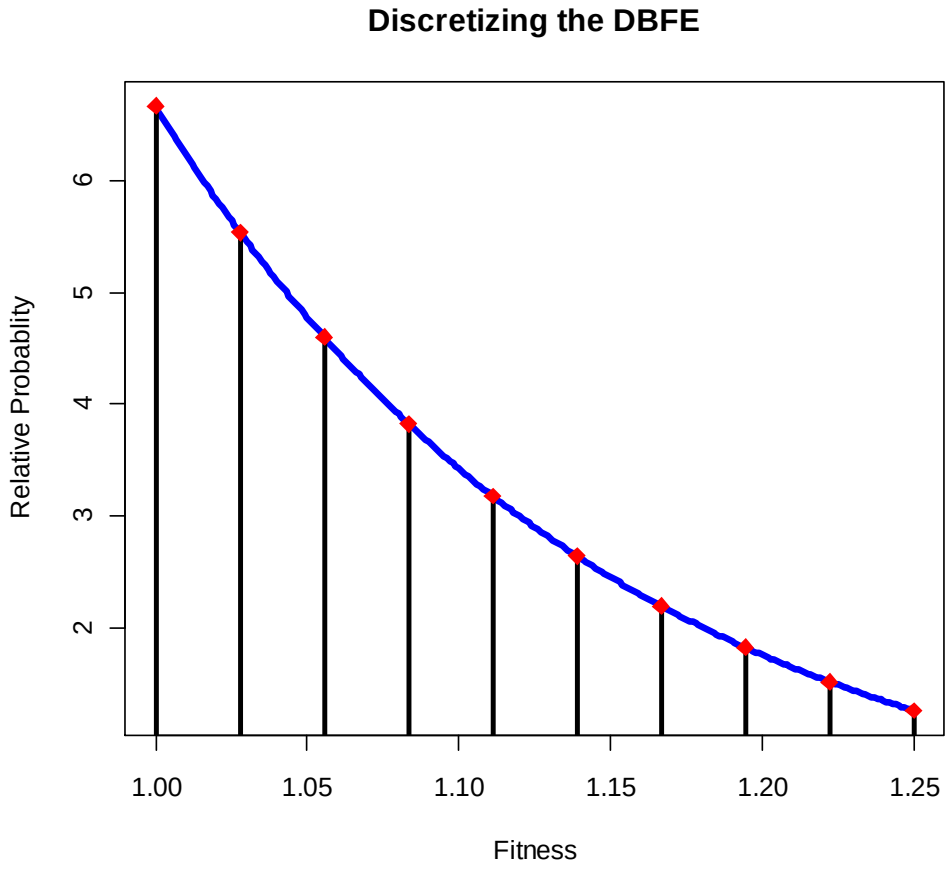


Figure 4.1 - A discretized model for the DBFE. Shown in blue is a continuous probability distribution that we wish to approximate. Samples from this probability distribution can be drawn from any continuous value along the number line. In contrast, samples from the discretized version of this distribution can only be drawn from the exact values indicated by the black lines in the figure. In both cases, however, the likelihood of a particular sample value is proportional to the value on the y-axis. In this manner, we can approximate any DBFE by using a discrete version of points.

To fit this model, we used the fitness values of the isolates obtained at the end of the evolution experiment (Figure 4.2). As many isolates appeared to have a neutral fitness and not have acquired a beneficial mutation, we first set the fitness value of any isolate to 1 if their estimated fitness in three replicated assays was not found to be significantly different from this value by a t-test. In addition to failing the t-test, these isolates almost surely did not have a beneficial mutation as their estimated selective advantage from competition assays was always +/- 0.5% and mutations with such small effects would not have had enough time to reach a significant frequency in the number of generations these populations were evolved for (as a simple example, ignoring mutation and drift, a variant with a 1% growth rate advantage present at the start of the experiment would deterministically still have a frequency less than $1/1000^{th}$ after 252 generations of evolution, even at the smallest population size). The remaining isolates were deemed to have a beneficial mutation, and we counted each isolate as an observation from the discrete fitness effect class, c_i, if the fitness value of that class was closest to that measured for the isolate.

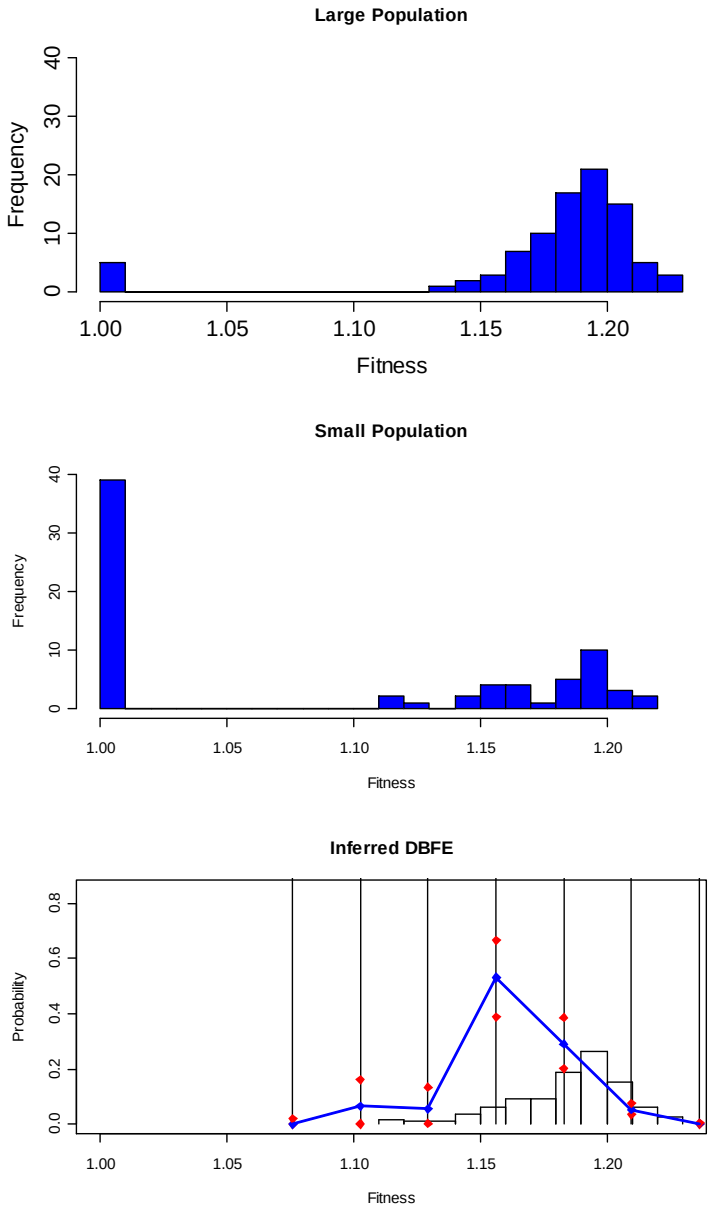


Figure 4.2- Distribution of fitness values for the isolates obtained from large and small populations at the end of the experiments. The top row shows the distribution in the large population, and the middle panel shows the distribution of fitness values in the small population. The bottom panel shows the DBFE for the discretized distribution as well as a

histogram of all values. The mean probability for that class is shown by the blue dot, while the upper and lower bounds of the 95% HPD interval is shown by the red dots. Note that the DBFE amongst the center of observed values is sensitive to how the DBFE is discretized, but the salient and robust conclusion is that values of intermediate fitness do not occur and must inevitably exist, if at all, at very low rates relative to those of the large effect mutations.

Having transformed the fitness data from the evolved isolates into counts from discrete fitness classes, we next inferred the model parameters. Note that in this model the total number of beneficial mutations that occur during the entire experiment is Poisson distributed, and conditioned on the total number of mutations that occurred, the number of mutations from each fitness class is distributed as a multinomial. In practice however, we only know the fitness of the isolates obtained at the end of the experiment, and not the total that occurred in all individuals at any point during the experiment. To circumvent this difficulty, we simply made the total number of mutations that occurred from each class in the entire experiment additional parameters in the model, and used Gibbs sampling to alternatingly update the model parameters U_b and $\boldsymbol{p}$, as well as the number of mutations that occurred throughout the experiment. While finding the posterior distribution for these parameters, we further made one additional simplification to the model to ease interpretability. We limited our inference about the DBFE only to regions of parameter space that this experimental design could provide information on. This allowed us to avoid presenting results that would be largely determined by the prior distributions we placed on the parameters. In particular, beneficial mutations with incredibly large effects but with infinitesimally small probabilities of occurring, or mutations that occur at reasonable frequencies but have incredibly small effects on fitness, can simply not be detected by our experiment; from a practical perspective, they are also not relevant

The discretized DBFE we inferred as well as the distribution of fitness values for isolates obtained at the end of the experiment are given in Fig 4.2. We estimated that beneficial mutations with effects at or above over 7.5% occur at a rate of 4×10^{-7} per generation (95% HPD $3.0 \times 10^{-7} - 5.2 \times 10^{-7}$). Importantly, the measurable DBFE appears to have its entire probability concentrated a large fitness effects, with a notable dearth of probability at intermediate values.

The beneficial mutation rate is high relative to the overall mutation rate

In order to compare the estimated beneficial mutation rate to the total per-cell beneficial mutation rate, we sequenced the genomes of six lines of *M. extorquens* AM1 that were propagated for 1,500 generations as part of a mutation accumulation experiment [38]. Illumina re-sequencing was performed for the end points of all six lines, and mutations were called in a manner previously described [16].

We observed a total of 25 mutations in these lines and estimated the per duplication mutation rate as 2.8×10^{-3} per genome per generation (95% CI 1.7×10^{-2} - 4.1×10^{-2}). The mutations were distributed as 12 IS element insertions, 11 SNPs and 2 genomic deletions. Comparing this estimate to our estimate of the beneficial mutation rate, we estimated that approximately 1 out of 10,000 mutations has a large beneficial effect in this environment.

Evolved Isolates are phenotypically convergent in both growth rate and stationary phase behavior

The convergence of the evolved isolates obtained at the end of the experiment on two different traits suggests that all harbor beneficial mutations that similarly affect cell physiology. The evolved isolates not only all had a ~20% improvement in growth rate (Fig. 4.2) but also had a

strikingly different growth behavior than the ancestor in stationary phase. The ancestral *M. extorquens* strains used in these experiments exhibited changes in OD during batch culture typical of many, if not most, bacteria. During growth the OD value exponentially increases, but after climaxing sharply decreases. This behavior is presumably due to the cells transitioning from a rod-shaped to egg-shaped morphology and becoming smaller, which occurs in *E. coli* and many other bacteria as they approach stationary phase [39]. However, the evolved isolates exhibited no decrease in OD values after reaching stationary phase, in contrast to the ancestral type (Fig. 4.3).

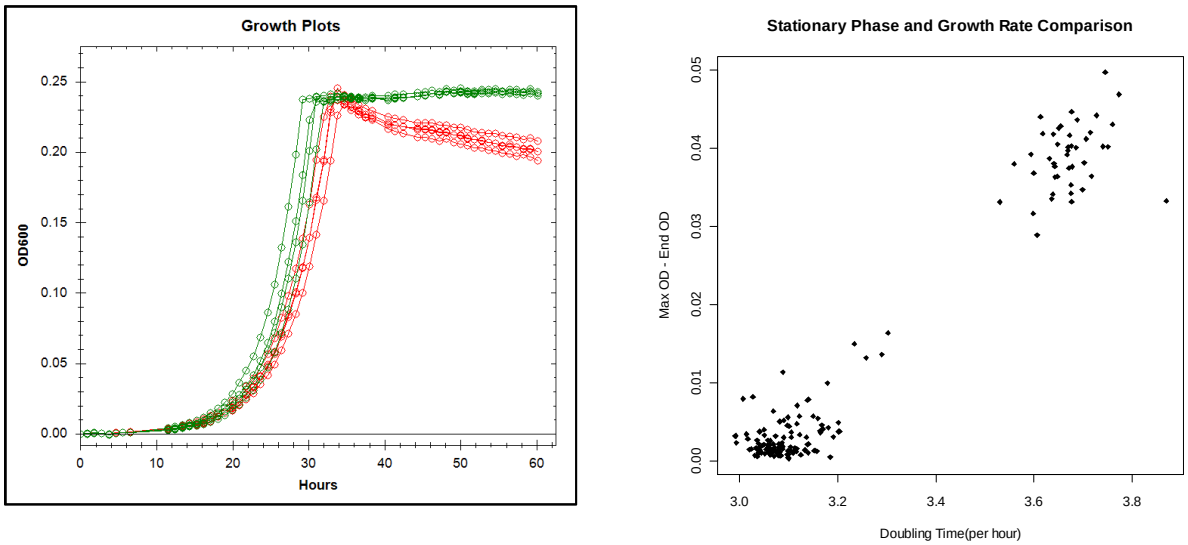


Figure 4.3 - Comparison of how the OD of changes in stationary phase. The left panel shows 8 example isolates from the evolution experiment, 4 in green that have a fast doubling time phenotype, and 4 in red that have a doubling time similar to the ancestor and have likely not acquired a beneficial mutation. The right panel shows a comparison of the doubling time of an isolate to the difference between the maximum OD reading final OD reading achieved at 60 hours.

Beneficial mutations were not substrate specific

We sought to determine if the adaptation we observed might be specific to growth on the substrate methylamine and so indicative of such a substrate specific and possibly simple mechanism by evaluating the growth rates of all evolved isolates from this experiment on the multi-carbon compound succinate. We found that the growth rate improvement on succinate was proportional to the improvement on methylamine, and so was not substrate specific (SI).

Posterior predictive checks are in agreement with the estimated DBFE and mutation rate

While inferring the DBFE and the associated mutation rate, we only used the fitness values of isolates obtained at the end of the experiment. We used other data collected during the experiment to verify the statistical model we used and to determine if the inferred parameters were in agreement with this additional data. We performed posterior predictive checks by simulating data from the parameter values at the mode of their posterior distribution, and then compared these simulations to the data that was excluded from the fitting process.

The large populations all have a high probability that a beneficial mutation escapes drift quickly, while the small populations must "wait" a randomly distributed amount of time for a beneficial mutation to appear; as a result, the simulations at the inferred parameter values showed that the large populations should all increase in fitness at approximately the same time, while the small populations should increase at more irregular times that are more evenly distributed (Fig. 4.4). We measured the growth rates of all the evolving populations at every transfer during this experiment, and recorded the times at which populations were in the middle of a selective sweep as a metric to check for this behavior. We found that in agreement with expectations the large

population treatments all had this occur after a relatively small number of generations centered around ~110 generations. In contrast, for the small population size treatment that did acquire a beneficial mutation, the number of generations required was more uniformly spread out in the small population size treatments, and even for the approximately half of the populations that did experience a sweep the average time was around generation 180 (Fig. 4.5).

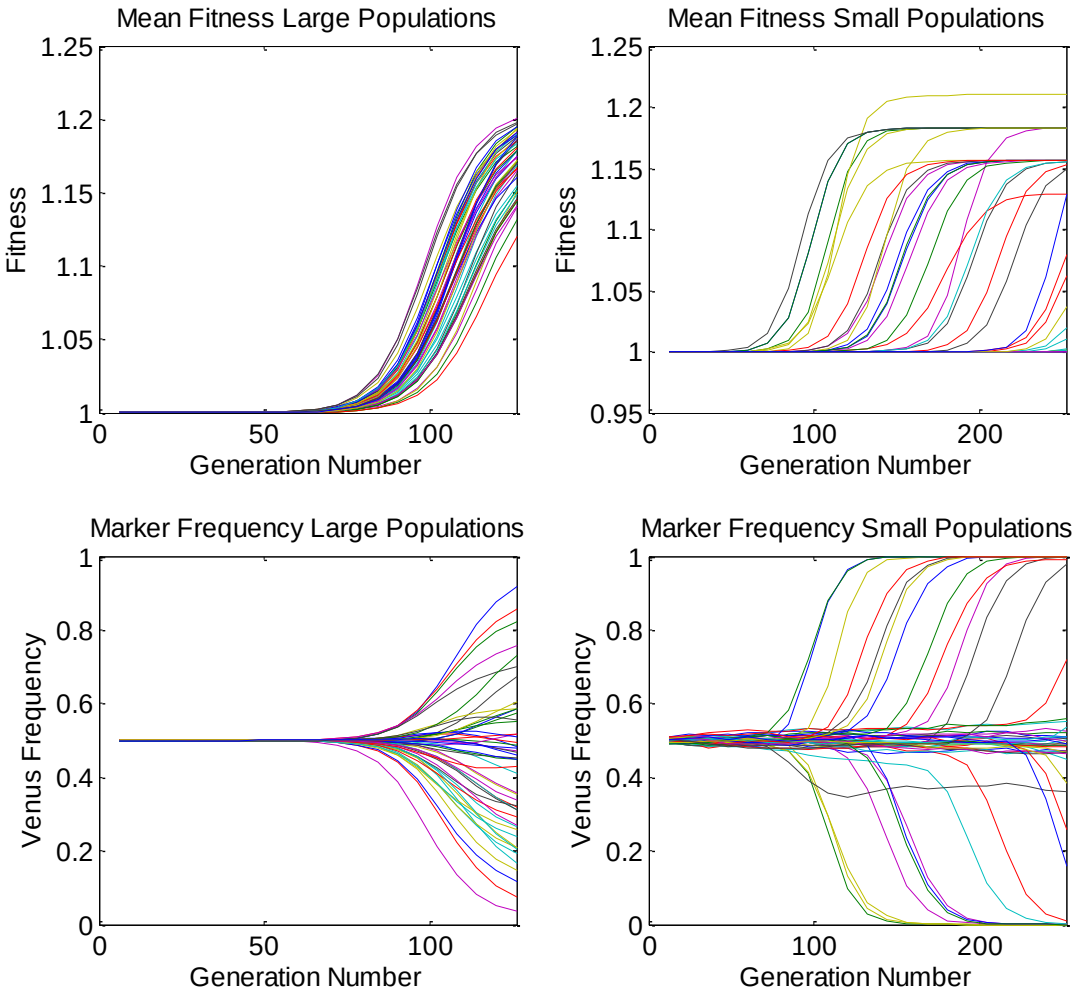


Figure 4.4 - Simulated data for 96 populations evolved at each population size using the parameter settings at the mode of their posterior distribution. Each line represents a different population.

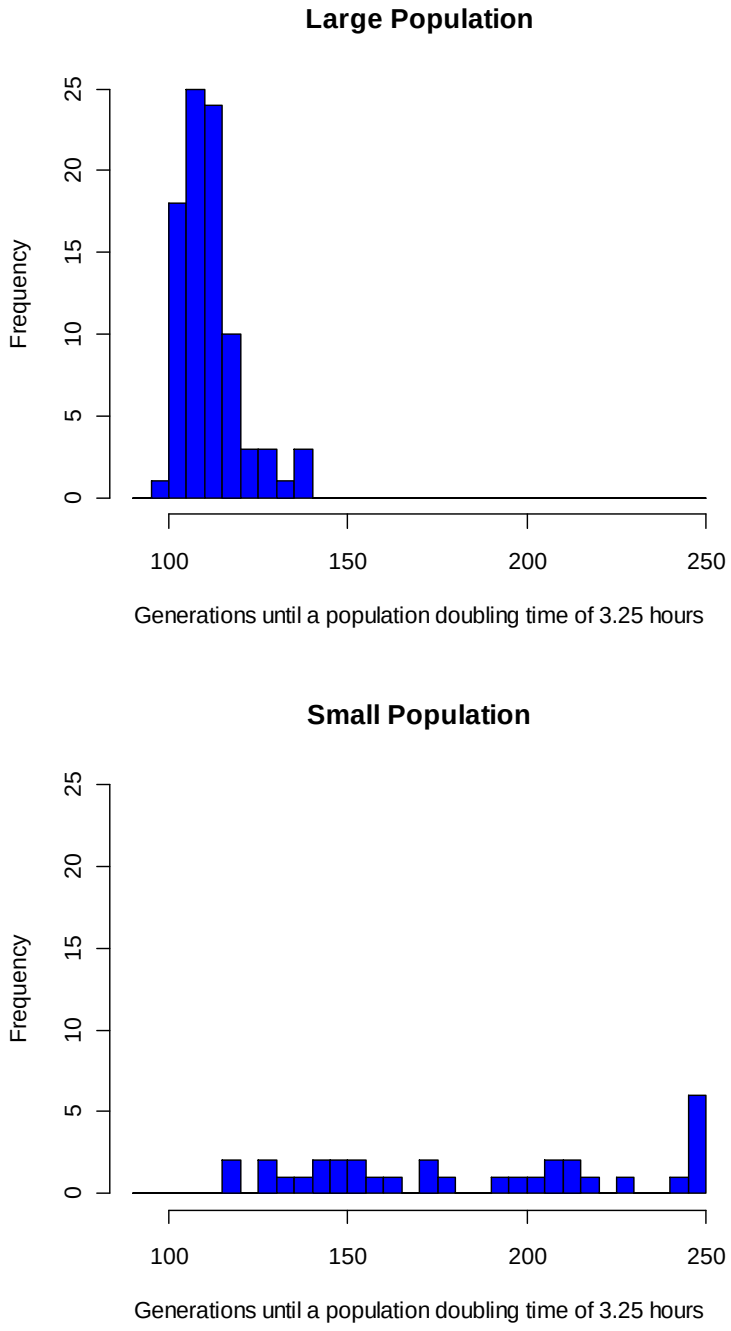


Figure 4.5 - Distribution of the number of generations required for the mean growth rate measured in a population to have a measured doubling time below 3.25 hours (the ancestor has a 3.5 hour time). In agreement with the simulations, the large population has much more consistent times than the small.

Further validation came from how the frequency of the different fluorescent markers changed over the course of the experiment. The large populations are likely to get multiple beneficial mutations in both fluorescent markers, creating clonal interference dynamics where, although the mean population fitness improves in a manner akin to periodic selection of a single mutation, it is quite likely for neither fluorescent protein to reach 100% frequency (Fig. 4.4). In contrast, mutations are rare enough in the small population that when a mutation appears it can usually completely sweep (Fig. 4.4). In agreement with this expectation, for each population we compared the estimated mean population growth rate at the end of the experiment to the deviation of the two fluorescent protein frequencies from the 50/50 value used at the start (Fig. 4.6). The large populations showed that the final marker frequency was uncorrelated with growth rate, while they were correlated in the small populations.

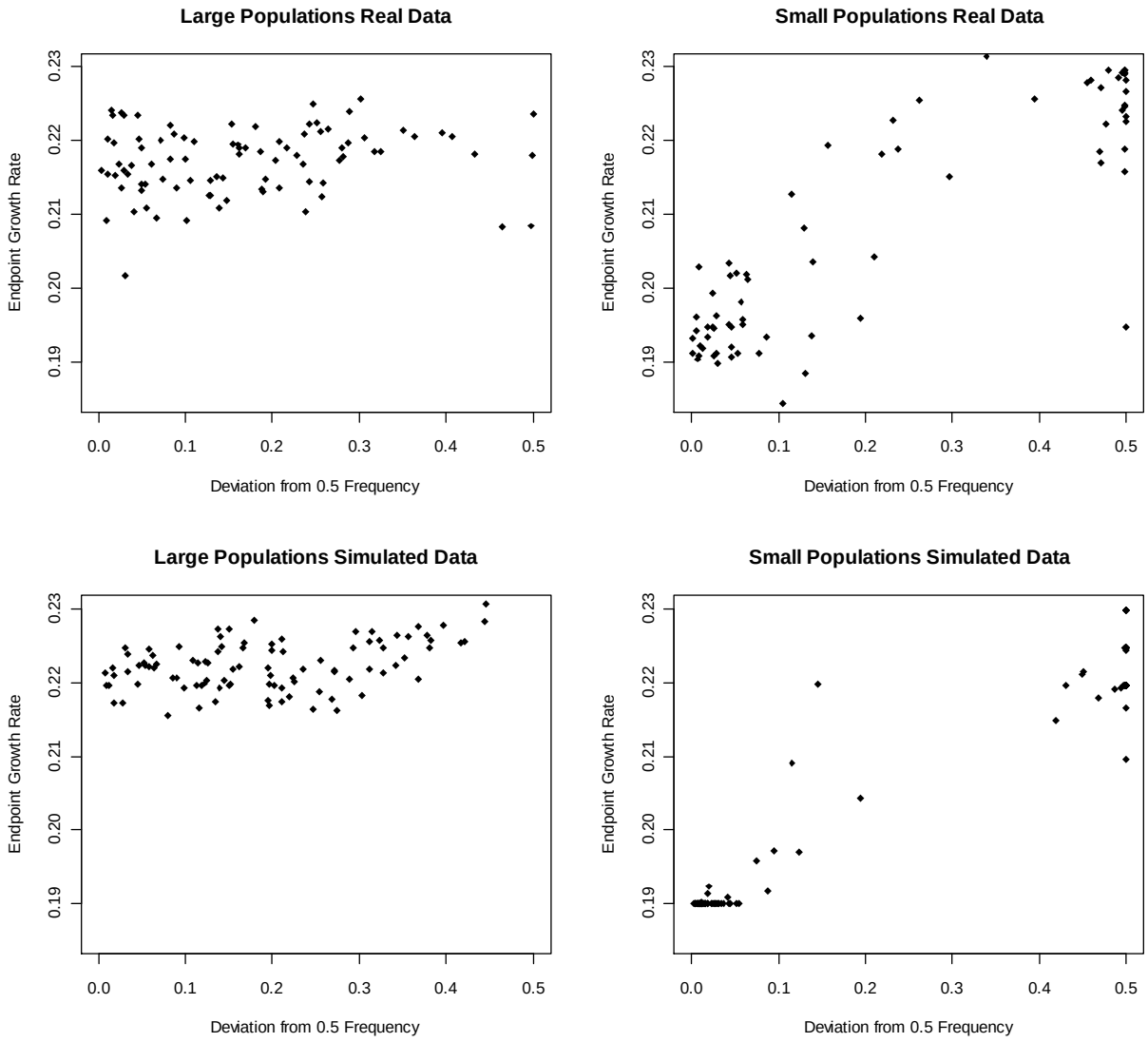


Figure 4.6 - Comparison of the mean estimated grow rate for evolved populations at the end of the experiment to the deviation of the fluorescent protein ratio from its starting value of 0.5, shown for both actual data and simulated data at both population sizes. The large populations show substantial improvement but not the predictable fixation of one marker-type, while these two things are correlated in the small population.

Discussion

Whether the mutations that underlie adaptation are many with small-effects or a few with large effects is one of the oldest debates in evolutionary biology, dating back to early correspondence between Huxley and Darwin [1,5,6,40,41]. Over the past three decades this debate has been advanced quantitatively by attempts to characterize the DBFE and the beneficial mutation rate, or particular aspects of them, in different organisms under different selective conditions [2,12,42,43,44,45,46,47,48]. One of the earliest assumptions about the shape of the DBFE in these studies is that it should reflect the intuition that small-effect mutations are more common than large effect mutations, and thus look like such parametric forms as an exponential distribution [1,9,10]. Unfortunately, although much of the best data to date about the nature of beneficial mutations has come from experimentally evolving microbes, most of this data has not allowed us to test this assumption. The number of replicate populations used have been too small and the population sizes have been too large to allow a complete and unbiased sampling of the DBFE. A single large population size only samples the rare mutants at the tail of the DBFE, while a single small population size would be too small to capture the truly rare variants, meaning that multiple population sizes are needed to effectively sample the DBFE.

In this project, we overcame these obstacles by evolving a large number of populations at two different sizes. Large-scale replicated evolution experiments for selection on growth rates can be difficult because they require a method to grow many populations under identical well-mixed conditions, so that the data produced will match the population genetic model, and also require controls for the contamination that may occur. We used a recently described culturing system

and uniquely barcoded ancestral strains to satisfy these requirements, creating an experimental design that should have allowed us to sample a diversity of selective effects from the DBFE.

However, despite our ability to detect smaller-effect mutations, we did not find them. We instead found that large-effect mutations occur at a high rate, and that the DBFE had a substantial 'hump' in it such that these large-effect mutations were more common than intermediate or small effect mutations. This statement is qualified because mutations with far smaller effects than the ones we observed could still occur. However, small effect mutations not only are less likely to stochastically escape drift, and so effectively occur at much lower rates, but they also take far longer than large-effect mutations to increase in frequency. This means that these large effect mutations will appear and fix before the small effect mutations can impact the adaptive dynamics even at population sizes much smaller than might previously have been thought.

We expected our large and small population treatments to recover qualitatively different isolates, but both evolved very high fitness types that showed a similar behavior in stationary phase. Although the mean selective advantage of isolates obtained from the large population treatment was higher, and these populations fixed evolved mutations faster (Fig. 2), the mutations that arose were still qualitatively very similar to those in the small population size treatments. One interesting result of this work is that it provides an alternative, and complimentary, explanation for the simplification of the DBFE that takes place when evolution occurs in large populations. Due to the effect of clonal interference, small effect mutations are not relevant to the dynamics of large populations; their dynamics can be approximated by a simple selective coefficient [21,22,25] which represents the narrow range of values that are both likely enough to occur and also be competitive. However, our results suggest that this effect results not only from these

competition dynamics, but is also a fundamental property of the DBFE at all population sizes. We found had no density for mutations with an intermediate fitness advantage.

Is this finding atypical or an anomalous result not applicable to other systems? Recent studies suggest that our result may be indicative of a common evolutionary pattern. Rokyta et. al recently examined the DBFE using beneficial mutations found from both a DNA and an RNA virus, and although viruses have much simpler genetic architectures than bacteria, they did not find support for the assumption that small effect sized mutations were more common [45]. Two studies that have examined the DBFE using antibiotics to facilitate sampling from it have also observed strong beneficial mutations rates and a DBFE with a hump [12,49], though in one case the shape of the DBFE found depended on the concentration of the antibiotic used, and other studies have also found support for an exponential model [47].

There is some discussion in the literature that a DBFE with many large-effect mutations is expected if the genotype being evolved is initially poorly adapted, whereas the DBFE for well-adapted genotypes should be expected to only have smaller-effects [13,49]. This notion has support from two observations that have consistently been made with experimental evolution. The first is that the rate of adaption of a population evolving under constant conditions typically slows through time [14,50] (though see [51] for a notable counter example). The second is that genotypes that have recently suffered a genetic insult that removes them from a high fitness state typically are able to recover that fitness, or compensate, very quickly by acquiring new mutations [52,53].

However, we would argue that without a well-defined historic or global optimum value for a trait, expecting different beneficial effect sizes based on the current fitness of an organism can be

tautological. For example, in a separate experiment, when *M. extorquens* was evolved for growth on succinate at large population sizes, the rate at which fitness increased did slow down as the experiment proceeded [54]; this could be considered evidence for diminished mutational effects with increased organismal fitness. However, even after 1,500 generations the doubling times of the evolved bacteria were over 2.5 hours, still far slower than other bacteria such as *E. coli*. In a global sense, the evolved bacteria were still very poorly adapted, yet the effect sizes of mutations had decreased. It is not the actual fitness of the organism, but rather the amount of time that the selection has occurred for that determines the effect size of the DBFE, and large effect sizes can be common when the selection is newly applied.

We have yet to determine the physiological basis underlying the physiologically convergent adaptation we observed, which could help to provide a mechanistic argument for the generality of these results. However, the observation that the evolved isolates we obtained not only have similar fitness effects but also share a distinctly different OD behavior in stationary phase suggests an interesting parallel with past evolution experiments that used microbes deeply diverged from *Methylobacterium*. As mentioned earlier, several experimental evolution studies using γ-proteobacteria under a variety of different conditions have frequently found mutations that impair the *rpoS* gene [27,29,30,55]. The product of this gene, σ^S, is consider a master regulator of stress response. Mutations to *rpoS* have many diverse effects on cell physiology [56,57,58] and as there are multiple mechanisms by which these mutations can be adaptive [56,57,59,60,61] they are often beneficial in both continuous culture and in stationary phase [56]. Strains deficient in *rpoS* do however mimic the evolved isolates from this study in that they do not shrink when they enter stationary phase [39]. Although *Methylobacterium*, being an α-

proteobacterium, does not have a *rpoS* gene to mutate, this convergent phenotype might be indicative of a physiological parallel between the types of mutations found.

Our estimate of the beneficial mutation rate, at 4×10^{-7} is very high, particularly since we only estimated the rate for mutations with fitness effects over 5%. That the mutation rate is this highly is almost certainly due to the presence of several mutational targets, all of which have an approximately similar selective effect and physiological response, but which vary enough to give the different fitness values seen here. One consequence of a beneficial mutation rate this high is that clonal interference and multiple mutation dynamics, which are sometimes referred to as soft sweeps in the eukaryotic literature, can become relevant at even small population sizes [62]. In addition, the strong-selection weak-mutation (SSWM) models of population dynamics [32,63], are unlikely to apply to as many scenarios as previously thought. Both of these conclusions have also found support recently for different reasons [62].

The most salient effect of such a high rate of large effect mutations though is that evolution is likely to proceed via relatively similar physiological steps at both large and small populations. With a diverse or monotonically decreasing BDFE one could expect less diversity of selective effects in large populations due to the filtering that takes place by clonal interference. However, with a BDFE that has high density only on large effects and a dearth of intermediate effects, even small populations will be physiologically very similar to their larger counterparts, and to the extent that these early mutations limit or influence later adaptation, one cannot necessarily expect to explore different adaptive trajectories by sampling different initial mutations at smaller population sizes.

References

1. Orr HA (2005) The genetic theory of adaptation: a brief history. Nature Reviews Genetics 6: 119-127.

2. Desai MM, Fisher DS, Murray AW (2007) The Speed of Evolution and Maintenance of Variation in Asexual Populations. Current Biology 17: 385-394.

3. Cooper TF (2007) Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli. PLoS Biology 5: e225.

4. Sniegowski PD, Gerrish PJ (2010) Beneficial mutations and the dynamics of adaptation in asexual populations. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 1255-1263.

5. Pritchard JK, Di Rienzo A (2010) Adaptation–not by sweeps alone. Nature Reviews Genetics 11: 665-667.

6. Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution.

7. Bell G, Gonzalez A (2009) Evolutionary rescue can prevent extinction following environmental change. Ecology Letters 12: 942-948.

8. Allen Orr H (1999) The evolutionary genetics of adaptation: a simulation study. Genetics Research 74: 207-214.

9. Beisel CJ, Rokyta DR, Wichman HA, Joyce P (2007) Testing the Extreme Value Domain of Attraction for Distributions of Beneficial Fitness Effects. Genetics 176: 2441.

10. Gillespie JH (1984) Molecular evolution over the mutational landscape. Evolution 38: 1116-1129.

11. Orr HA, Coyne JA (1992) The Genetics of Adaptation: A Reassessment. American Naturalist 140: 725.

12. McDonald MJ, Cooper TF, Beaumont HJE, Rainey PB (2011) The distribution of fitness effects of new beneficial mutations in Pseudomonas fluorescens. Biology Letters 7: 98-100.

13. Barrett RDH, Craig MacLean R, Bell G (2006) Mutations of intermediate effect are responsible for adaptation in evolving Pseudomonas fluorescens populations. Biology Letters 2: 236-238.

14. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. Nature Reviews Genetics 4: 457-469.

15. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, et al. (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461: 1243-1247.

16. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332: 1190.

17. Dettman JR, Rodrigue N, Melnyk AH, Wong A, Bailey SF, et al. (2012) Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. Molecular ecology.

18. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF (2011) Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332: 1193-1196.

19. Gerrish PJ, Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. Genetica 102-3: 127-144.

20. Gumbel EJ (2004) Statistics of extremes: Dover Publications.

21. Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM (2012) Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proceedings of the National Academy of Sciences 109: 4950-4955.

22. Hegreness M, Shoresh N, Hartl D, Kishony R (2006) An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311: 1615-1617.

23. Marx CJ (2011) Evolution as an experimental tool in microbiology:'Bacterium, improve thyself!'. Environmental Microbiology Reports 3: 12-14.

24. Portnoy VA, Bezdan D, Zengler K (2011) Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. Current Opinion in Biotechnology 22: 590-594.

25. Atsumi S, Wu TY, Machado IMP, Huang WC, Chen PY, et al. (2010) Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Molecular Systems Biology 6: 449.

26. Minty JJ, Lesnefsky AA, Lin F, Chen Y, Zaroff TA, et al. (2011) Evolution combined with genomic study elucidates genetic bases of isobutanol tolerance in Escherichia coli. Microbial Cell Factories 10: 18.

27. Conrad TM, Joyce AR, Applebee MK, Barrett CL, Xie B, et al. (2009) Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. Genome Biology 10: R118.

28. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, et al. (2006) Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nature Genetics 38: 1406-1412.

29. Kinnersley MA, Holben WE, Rosenzweig F (2009) E Unibus Plurum: genomic analysis of an experimentally evolved polymorphism in Escherichia coli. PLoS Genetics 5: e1000713.

30. Maharjan R, Zhou Z, Ren Y, Li Y, Gaffé J, et al. (2010) Genomic identification of a novel mutation in hfq that provides multiple benefits in evolving glucose-limited populations of Escherichia coli. Journal of bacteriology 192: 4517-4521.

31. Burch CL, Chao L (1999) Evolution by small steps and rugged landscapes in the RNA virus φ6. Genetics 151: 921-927.

32. Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science 312: 111-114.

33. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in Methylobacterium extorquens AM1 from a genomic point of view. Journal of bacteriology 185: 2980-2987.

34. Wahl LM, Gerrish PJ (2001) The probability that beneficial mutations are lost in populations with periodic bottlenecks. Evolution 55: 2606-2610.

35. Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. American Naturalist: 1315-1341.

36. Delaney NF, Kaczmarek ME, Ward LM, Swanson PK, Lee MC, et al. (2012) Development of an optimized medium, strain and high-throughput culturing methods for *Methylobacterium extorquens*. PLoS One Submitted.

37. Lang GI, Botstein D, Desai MM (2011) Genetic variation and the fate of beneficial mutations in asexual populations. Genetics 188: 647-661.

38. Lee MC, Marx CJ (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genetics 8: e1002651.

39. Lange R, Hengge-Aronis R (1991) Growth phase-regulated expression of bolA and morphology of stationary-phase Escherichia coli cells are controlled by the novel sigma factor sigma S. Journal of Bacteriology 173: 4474-4481.

40. Bell G. The oligogenic view of adaptation; 2009. Cold Spring Harbor Laboratory Press. pp. 139-144.

41. Provine WB (1987) The origins of theoretical population genetics. Chicago: University of Chicago Press. xi, 201 p.

42. Imhof M, Schlotterer C (2001) Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. PNAS 98: 1113-1117.

43. Joseph SB, Hall DW (2004) Spontaneous Mutations in Diploid *Saccharomyces cerevisiae* More Beneficial Than Expected. Genetics 168: 1817-1825.

44. Kassen R, Bataillon T (2006) Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nature Genetics 38: 484-488.

45. Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, et al. (2008) Beneficial fitness effects are not exponential for two viruses. Journal of Molecular Evolution 67: 368-376.

46. Sanjuan R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. Proceedings of the National Academy of Sciences 101: 15376.

47. Schenk MF, Szendro IG, Krug J, de Visser JAGM (2012) Quantifying the Adaptive Potential of an Antibiotic Resistance Enzyme. PLoS genetics 8: e1002783.

48. Schoustra SE, Bataillon T, Gifford DR, Kassen R (2009) The properties of adaptive walks in evolving populations of fungus. PLoS biology 7: e1000250.

49. MacLean RC, Buckling A (2009) The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS genetics 5: e1000406.

50. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332: 1190-1192.

51. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proceedings of the National Academy of Sciences 105: 7899-7906.

52. Poon A, Chao L (2005) The rate of compensatory mutation in the DNA bacteriophage φX174. Genetics 170: 989-999.

53. Poon A, Davis BH, Chao L (2005) The Coupon Collector and the Suppressor Mutation Estimating the Number of Compensatory Mutations by Maximum Likelihood. Genetics 170: 1323-1332.

54. Lee MC, Chou HH, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of *methylobacterium* to distinct growth substrates. Evolution 63: 2816-2830.

55. Wang L, Spira B, Zhou Z, Feng L, Maharjan RP, et al. (2010) Divergence involving global regulatory gene mutations in an Escherichia coli population evolving under phosphate limitation. Genome Biology and Evolution 2: 478.

56. Ferenci T (2005) Maintaining a healthy SPANC balance through regulatory and mutational adaptation. Molecular microbiology 57: 1-8.

57. King T, Seeto S, Ferenci T (2006) Genotype-by-environment interactions influencing the emergence of rpoS mutations in Escherichia coli populations. Genetics 172: 2071-2079.

58. Patten C, Kirchhof M, Schertzberg M, Morton R, Schellhorn H (2004) Microarray analysis of RpoS-mediated gene expression in Escherichia coli K-12. Molecular Genetics and Genomics 272: 580-591.

59. Ihssen J, Egli T (2004) Specific growth rate and not cell density controls the general stress response in *Escherichia coli*. Microbiology 150: 1637-1648.

60. Stoebel DM, Hokamp K, Last MS, Dorman CJ (2009) Compensatory evolution of gene regulation in response to stress by *Escherichia coli* lacking RpoS. PLoS Genetics 5: e1000671.

61. Zambrano MM, Siegele DA, Almirón M, Tormo A, Kolter R (1993) Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures. Science (New York, NY) 259: 1757.

62. Karasov T, Messer PW, Petrov DA (2010) Evidence that Adaptation in *Drosophila* Is Not Limited by Mutation at Single Sites. PLoS Genetics 6: e1000924.

63. Gillespie JH (1991) The Causes of Molecular Evolution: Oxford University Press, USA.

Chapter 5

Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*

A study of evolutionary rates and processes following a host shift.

Due to formatting requirements, this work is reprinted in the back matter.

Reprinted from PLoS Genetics

Delaney, Nigel F., et al. "Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, *Mycoplasma gallisepticum*." *PLoS genetics* 8.2 (2012): e1002511.

Chapter 1 Back Matter

Clarity: An Open-Source Manager for Laboratory Automation

A description of the software created to perform the experiments in this dissertation.

Reprinted from the Journal of Laboratory Automation.

Delaney, N. F., Echenique, J. I. R., & Marx, C. J. (2012). Clarity an Open-Source Manager for Laboratory Automation. *Journal of Laboratory Automation*.

Clarity: An Open-Source Manager for Laboratory Automation

Nigel F. Delaney[1], José I. Rojas Echenique[1], and Christopher J. Marx[1,2]

Abstract

Software to manage automated laboratories, when interfaced with hardware instruments, gives users a way to specify experimental protocols and schedule activities to avoid hardware conflicts. In addition to these basics, modern laboratories need software that can run multiple different protocols in parallel and that can be easily extended to interface with a constantly growing diversity of techniques and instruments. We present Clarity, a laboratory automation manager that is hardware agnostic, portable, extensible, and open source. Clarity provides critical features including remote monitoring, robust error reporting by phone or email, and full state recovery in the event of a system crash. We discuss the basic organization of Clarity, demonstrate an example of its implementation for the automated analysis of bacterial growth, and describe how the program can be extended to manage new hardware. Clarity is mature, well documented, actively developed, written in C# for the Common Language Infrastructure, and is free and open-source software. These advantages set Clarity apart from currently available laboratory automation programs. The source code and documentation for Clarity is available at http://code.google.com/p/osla/.

Keywords

laboratory information managements systems (LIMS), informatics and software, robotics and instrumentation engineering, HTS high-throughput screening, automated biology

Introduction

Robotic automation is revolutionizing research in fields including clinical science,[1] genomics,[2] and systems biology.[3,4] Automated laboratories can produce better, more consistent data; can have lower operating costs; and can be scaled up easily. As more laboratories begin to embrace the benefits of automation, the programs that are used to manage laboratory instruments will have to confront the needs of a new and more diverse group of users.

Software to manage automated laboratories has to interface with hardware instruments, give users a way to describe the activities that make up experimental protocols, and schedule these activities in a way that avoids hardware conflicts. In addition to these basic requirements, modern laboratories need software that can run multiple different protocols in parallel. In a laboratory with independent investigators who share common equipment, software has to be able to schedule parallel protocols on demand without interrupting running protocols. It is also essential for software to be easily extensible so that it can adapt to a constantly growing diversity of techniques and instruments.

Here we present Clarity, a laboratory automation manager designed to meet the challenges of modern laboratory automation. Clarity is hardware agnostic, portable, extensible, and open source. Furthermore, it provides critical features that include remote monitoring, robust error reporting by phone or email, and full state recovery in the event of a system crash. We present the basic organization of Clarity, an example of its implementation for the automated analysis of bacterial growth, and a description of how the program can be extended with new instrument interfaces and graphical user interfaces.

General Attributes of the Organization of Clarity

Hardware and Task Management

The automation of even the most rudimentary laboratory procedures often requires the orchestration of multiple specialized instruments. Instruments usually serve different functions and communicate with the computer in different

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA
[2]Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, Massachusetts, USA

Corresponding Author:
Christopher J. Marx, Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA, 02143, USA
Email: cmarx@oeb.harvard.edu

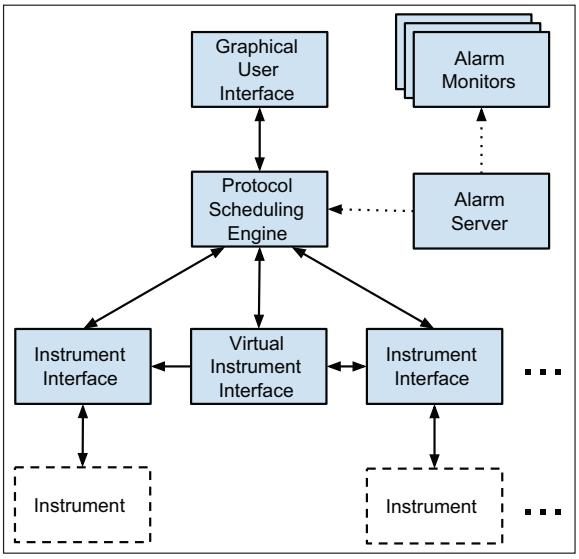


Figure 1. A diagram of Clarity's important components and their organization. Solid arrows indicate direct communication (e.g., the protocol-scheduling engine calls methods in the instrument interfaces that define activities); dashed arrows indicate connections over Internet protocols.

```
<Protocol>
  <ProtocolName>Example Protocol</ProtocolName>
  <ErrorEmailAddress>address@provider</ErrorEmailAddress>
  <ErrorPhoneNumber>5555555555</ErrorPhoneNumber>
  <Variables />
  <Instructions>
     <Instruction InstType="Clarity.StaticProtocolItem">
        <InstrumentName>RoboticArm</InstrumentName>
        <MethodName>MoveToPosition</MethodName>
        <Parameters>
           <Parameter Type="System.Int32">2</Parameter>
        </Parameters>
     </Instruction>
     <!- More instructions...  ->
  </Instructions>
</Protocol>
```

Figure 2. A simplified example of an XML protocol file.

ways. However, instruments are unified by their purpose: to perform meaningful activities during an experiment.

In Clarity, each instrument is associated with a dedicated instrument interface (**Fig. 1**). The instrument interface handles the low-level communication between software and hardware and defines a set of meaningful activities that the instrument can perform during the course of an experiment. The interface to a robotic arm, for example, might define an activity to move the arm to a specific location or lift a microtiter plate. This activity could then appear as a step in the protocol for an experiment. The instrument interface would be responsible for loading the positions of the incubator and spectrometer from a configuration file, for instructing the robotic arm to power the right motors in the right sequence, and for reporting back to Clarity in the event of any hardware errors.

Clarity also supports virtual instruments. Virtual instruments do not correspond to hardware; instead, they are meant to perform completely computational activities during experiments. They can be used to write log or data files, to organize hardware instrument interfaces for multi-instrument activities, or to monitor data output to make decisions about the course of an experiment.

Protocol Execution

Protocols define the activities that make up an experiment.[5] A simple protocol can consist of a list of activities and the times at which they should be performed. More complex protocols can incorporate control flow elements such as loops and conditional statements. Conditions are evaluated by virtual instruments and can therefore depend on anything that the program has access to, including data files, protocol details, and program state. Protocols can be written using a simple protocol description language based on XML, the eXtensible Markup Language.[6] **Figure 2** contains a simplified snippet from a typical protocol file. Alternatively, Clarity is equipped with a graphical protocol editor. Using the protocol editor, the user simply chooses activities from an interactive list and arranges them to specify her protocol. After providing an email address and phone number, for error reporting, the user can save the protocol to an XML file and use Clarity to execute it.

Clarity's scheduling engine keeps track of running protocols and uses the instrument interfaces to call the right activities at the right times. Scheduling is complicated by conflicts, which can arise when multiple users run protocols at the same time and on the same instruments. To resolve conflicts, the scheduler runs a simple and flexible algorithm: (1) When an activity finishes, it activates the scheduler. (2) The scheduler inspects the uncompleted activities of the remaining protocols and identifies the activity with the earliest prescribed time. (3) If that time is in the future, Clarity waits; otherwise, it executes the activity immediately. This algorithm is not ideal for procedures that are extremely time sensitive, but it is easy to run dynamically, meaning that new protocols can be added at any time without stopping the execution of running protocols.

Clarity's basic scheduling algorithm is designed to be easy to understand and to simply avoid any resource conflicts between different protocols. It runs multiple protocols in parallel by alternating which protocol is running serially at any moment, giving exclusive control of the entire system to one executing protocol and passing control of the system to another protocol (or context switching) only after the currently executing protocol has stopped using the system resources and returned them to a ready state. However,

Clarity, being open source, can also implement more complex scheduling schemes and run protocols in a truly concurrent manner that uses multiple instruments simultaneously for different tasks. Such parallel execution can, however, create resource conflicts, race conditions, deadlocks, and other problems. This is particularly difficult in the laboratory automation context, because which state a protocol is suspended in can be very important. For instance, we might not want to remove an item from an incubator and place it in a liquid handler if it will be some time before the liquid handler finishes its current task and is available to do the next step. For this reason, specifying a framework that optimally handles all possible concurrency issues, ensures that all the available instrument interfaces can provide enough information for the framework to appropriately make decisions (such as the time required to execute instructions), and does not introduce too much complexity to new users is difficult.

Instead of providing a general solution, Clarity's design assumes that the parallel execution problem for any specific usage scenario will be easier to solve by writing code for an idiosyncratic implementation than specifying what the problem for a general framework will be. Clarity provides the tools for a user to code relatively easily a more truly concurrent scheduler. Clarity allows users to implement concurrent operations through the use of virtual instrument classes. Instrument interfaces and virtual instrument objects can have direct access to the scheduler, the instruments, and all the loaded protocols. A user can simply write a virtual instrument that examines the entire system state and adjust the protocols and their execution order accordingly. Clarity also allows virtual instruments to handle events generated by its engine based on instrument processes, allowing them to respond to the actions taken by different protocols. A walk-through tutorial showing how to create concurrent or dynamic protocols and allow users to write more sophisticated scheduling algorithms to replace Clarity's simple scheduler is part of Clarity's online documentation. Clarity's modular organization, open-source license, thorough documentation, and community support make it especially amenable to this kind of customization. However, we emphasize that it will be the responsibility of anyone implementing such custom scheduling operations to ensure that the problems that can arise in parallel computing, such as deadlocks and race conditions, do not occur.

Error Reporting and Recovery

An unavoidable aspect of laboratory automation is that instruments can malfunction in the course of protocol execution. Some instrument interfaces can recognize and recover from common errors without user intervention. When an instrument interface encounters an error that it cannot handle, Clarity logs the error and immediately stops protocol execution. At this point, Clarity tries to alert the owners of the affected protocols about the error. Based on the user's preference, Clarity can send emails or text messages with detailed information about the probable causes of the error. Clarity can also call users' phone numbers to alert them at any hour of the day and night.

Error reporting is handled jointly by Clarity and a remote alarm server. The alarm server runs on a separate computer and exchanges information with Clarity over the Internet. This ensures that users continue to get error reports in the event that one of the computers malfunctions. The alarm server also lets users monitor running protocols. Users can install a monitoring program on a home computer and use it to connect to the alarm server over the Internet. The monitor displays upcoming activities, protocol information, and video from user-installed cameras.

Once notified, a user can often resolve problems remotely. Each instrument interface defines methods to reinitialize its associated instrument. The user can use Clarity's logs or the video cameras to determine whether an instrument needs to be reinitialized. If so, the user can activate the right recovery method from Clarity's graphical user interface.

Clarity always maintains a backup of the program's state: the list of running protocols and the list of activities that have yet to be performed. Before and after executing any activity, Clarity updates this backup. This ensures that when errors occur, there is a record of the program's state that can be used to rescue experiments. Once the problem is fixed, the backup can be loaded into a new instance of Clarity to continue running experiments as before.

Graphical Interface

Users interact with Clarity's components—hardware and virtual instrument interfaces, the protocol-scheduling engine, and the remote alarm server—through a graphical interface (**Fig. 3**). The main menu allows users to load and save program states, load protocols, and manage the remote alarm server. The body of the interface is organized into tabs, making it easy for users to add location-specific features. The main tab displays running protocols, instrument statuses, a log of errors, and controls to start and stop protocol execution. The error recovery tab provides methods to recover and reinitialize connected instruments. Additional tabs can be implemented to control specific instruments or to carry out location specific tasks.

Clarity's graphical interface updates itself automatically to accommodate new instrument interfaces. For example, the graphical protocol editor automatically includes activities from new or custom instruments. When it starts, the protocol editor generates its list of activities dynamically by inspecting all available instrument interfaces. The error recovery tab is also generated at run time. Clarity's self-updating user interface means that users can create custom

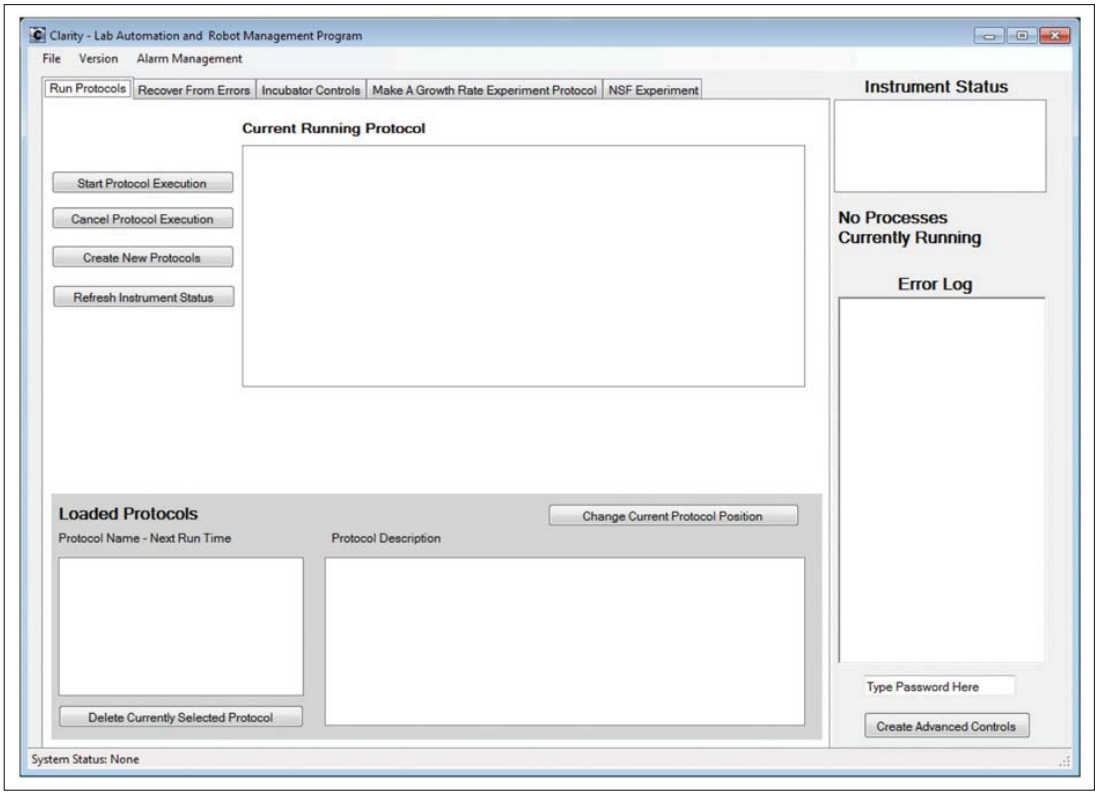


Figure 3. The main tab of Clarity's graphical use interface. This is where users can control protocol execution. The main tab also displays information about the currently running protocols, all scheduled activities, the statuses of hardware instruments, and an error log.

instrument interfaces without worrying about integrating them with the rest of the program.

Clarity in Action: Implementation for Automated Analysis of Bacterial Growth

To demonstrate a typical use case, we describe our laboratory's use of Clarity to manage a series of instruments for automated monitoring of bacterial growth (**Fig. 4**). A major effort in our group is to evolve replicate populations of one or more bacterial species in the laboratory as a means to study the physiological basis of adaptation.[7] Given that single experiments can involve hundreds of replicate populations, we maintain populations in 48-well microtiter plates that are stacked on an arbitrary-access, shaking tower that holds up to 38 plates. To maintain optimal growth of our study organism, *Methylobacterium extorquens*, we house the shaking tower, as well as the rest of the system, in a temperature-controlled environmental room at 30 °C and use a commercial humidifier to augment the humidity to ~75% relative humidity to minimize evaporation. Under these conditions, the primary component of fitness is the exponential growth rate of the culture.[8] By using a multiwell plate reader to take optical density readings over multiple days, we can assay the growth rate of nearly 2000 strains concurrently.

Users load a 48-well plate—already containing the needed media and cultures—onto the shaking tower. Using Clarity's graphical interface, the user specifies parameters of the growth curve protocol (e.g., number of measurements), selects the correct position on the tower, enters a file name for the data, and provides email addresses and phone numbers for error reporting. At this point, the user can also choose to apply the protocol to multiple plates, specifying their positions on the shaking tower. Then, the user initiates the protocol. A video demonstrating an automated measurement of optical density of a 48-well plate is available at http://www.evolvedmicrobe.com/LabAutomation.html. Briefly, the protocol proceeds as follows: (1) The

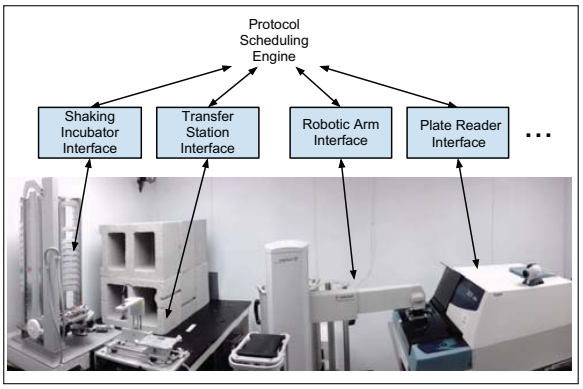


Figure 4. Our hardware configuration for tracing bacterial growth curves in an automated manner.

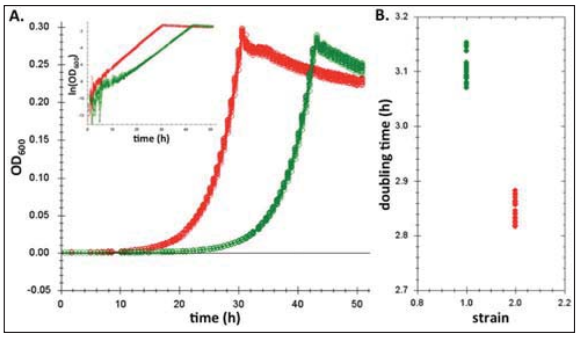


Figure 5. An example set of bacterial growth curves measured by Clarity. (**A**) The replicates represent a comparison of the optical density measured for over 50 h for two isolates (red and green; 18 replicates of each) of a strain of *Methylobacterium extorquens* AM1 evolved during an evolution experiment from the same ancestor. Note that in the inset, the log-linear increase in density indicates the remarkable constancy of per-capita growth throughout these conditions. (**B**) Analysis of the doubling times of the two strains indicates the precision by which we can estimate this exponential rate.

robotic spatula removes the microtiter plate from the specified position and places it on the transfer station. (2) A vacuum suction cup raises the plate's lid, and the plate base is moved toward the robotic arm. (3) The arm grasps the plate base, swings into position above the multiwell plate reader, and lowers the plate onto that instrument's loading platform. (4) The plate is lowered into the plate reader, and optical density readings are taken. A virtual instrument records the reading and all the relevant metadata to a spreadsheet. (5) Activities 1 to 4 are repeated in reverse to return the plate to the transfer station, replace the lid, and load the plate back into the tower. Each of the above activities is specified by the protocol, initiated by the scheduler, mediated through an instrument interface, and carried out by a particular instrument.

The data produced by Clarity are not only high throughput but also high quality. Because the 48-well plates allow for effective mixing and aeration—and due to our efforts to optimize the growth medium for *Methylobacterium*[9]—we routinely observe per-capita growth that is incredibly stable throughout exponential phase (**Fig. 5A**). Using a software package we have developed,[10] we can reliably measure small differences in growth rates (**Fig. 5B**). Because these small differences can have dramatic evolutionary consequences, it is crucial that our data be as high quality as possible. The required precision, scale, and extended timeline of our work preclude performing these experiments in the absence of automation.

Customizing Clarity

Clarity is easy to extend or customize with new instrument interfaces. Custom instrument interfaces can be written to control new hardware instruments or to carry out computational activities by implementing virtual instruments. Instrument interfaces can be written in any language that conforms to the Common Language Infrastructure.[11] Because most modern programming languages have a Common Language Infrastructure implementation, almost anyone with some programming experience can write a custom instrument. Furthermore, instrument interfaces are implemented as independent libraries that are loaded dynamically; this means that they can be included in the program without recompiling Clarity.

Clarity's online documentation[12] includes a tutorial on implementing a new instrument interface. There we demonstrate how to implement a virtual instrument to send email updates on the progress of running protocols. Every instrument interface will be different in the way that it handles communication with a hardware instrument—or email server—but all conform to a standard way of communicating with Clarity. Basically, every instrument interface needs to inherit from Clarity's BaseInstrumentClass class. This ensures that all interfaces define an instrument status flag, a recovery method, and a method to release system resources. In addition, BaseInstrumentClass implements a method to initialize location specific variables (e.g., network details) using XML configuration files. The class membership also serves to identify instrument interface classes. When the program starts, it looks for members of BaseInstrumentClass to include in the list of available instruments.

It is also possible to customize Clarity's graphical user interface. For example, we make extensive use of a tab that facilitates the design of growth curve protocols. The online documentation includes a template graphical user interface that can be tweaked and customized easily. Alternatively, because protocols reside in XML files, one can write a standalone protocol-generating application that would not

need to interface with Clarity directly. Instructions for customizing the user interface and sample interface templates are available in Clarity's documentation. Clarity is currently designed to manage lab instruments and execute instructions with them; however, it does not include a specific framework to manage the data generated by these experiments. A recent open-source database schema was published, AutoLabDB,[13] that would be useful for this endeavor, and future development will likely focus on implementing database interactions through virtual instruments in Clarity.

Conclusions

Clarity is mature, well-documented, and actively developed software for managing laboratory automation. We manage two automated laboratories with Clarity and plan on continuing to release bug fixes and new features. Because the success of a software project depends on the availability of quality support for new users, we maintain up-to-date documentation for Clarity online and are available to answer questions on a dedicated email list.[12] Clarity is written in the C# language and runs on the Common Language Infrastructure.[11] This means that the program can run on most software and hardware platforms (including Windows, OS X, and GNU/Linux operating systems). It also means that Clarity can be developed in almost any programming language.

Although Clarity currently has support for only a few instrument types, if a common protocol existed for interfacing with devices, it would be possible to have Clarity generically interact with any device implementing that protocol. Recently, the Standardization in Lab Automation (SiLA) consortium has created standards for device control interfaces that expose devices as web services and communicates with them using the simple object access protocol.[14] Devices using such a protocol could readily be made available for use in Clarity, and this is an active goal of the development team, but we are hindered only because we do not have access to any devices implementing the standardized protocols.

Clarity is open source under a free software license. This ensures that experiments performed by Clarity are as reproducible as possible because anyone can inspect the source code to determine how the program works and because experiments are specified in full detail by sharable XML protocols. It also ensures that Clarity remains flexible because anyone can modify the source code to fit particular needs. Most importantly, the open source paradigm means that every contribution to Clarity benefits the whole community of users. These advantages set Clarity apart from the other currently available laboratory automation programs.[15–17]

Acknowledgments

Co-authors N. F. Delaney and J. I. Rojas Echenique contributed equally to this work. We thank D. G. Robinson for his useful comments on this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the NIH (GM078209) and NSF (DEB-0845893).

References

1. Boyd, J. Robotic Laboratory Automation. *Science* **2002**, *295*, 517–518.
2. Mardis, E. R. A Decade's Perspective on DNA Sequencing Technology. *Nature* **2011**, *470*, 198–203.
3. King, R. D.; Rowland, J.; Oliver, S. G.; Young, M.; Aubrey, W.; Byrne, E.; Liakata, M.; Markham, M.; Pir, P.; Soldatova, L. N.; Sparkes, A.; Whelan, K. E.; Clare, A. The Automation of Science. *Science* **2009**, *324*, 85–89.
4. Zimmermann, H. F.; Degussa, J. R. A Fully Automated Robotic System for High Throughput Fermentation. *J. Lab. Autom*. **2006**, *11*, 134–137.
5. Schäfer, R. Concepts for Dynamic Scheduling in the Laboratory. *J. Lab. Autom*. **2004**, *9*, 382–397.
6. Bray, T.; Paoli, J.; Sperberg-McQueen, C. M.; Maler, E.; Yergeau, F. Extensible Markup Language (XML). **2008**, http://www.w3.org/XML/Core/#Publications. Accessed June 2012.
7. Lee, M.-C.; Chou, H.-H.; Marx, C. J. Asymmetric, Bimodal Trade-offs during Adaptation of *Methylobacterium* to Distinct Growth Substrates. *Evolution* **2009**, *63*, 2816–2830.
8. Chou, H.-H.; Chiu, H.-C.; Delaney, N. F.; Segrè, D.; Marx, C. J. Diminishing Returns Epistasis among Beneficial Mutations Decelerates Adaptation. *Science* **2011**, *332*, 1190–1192.
9. Delaney, N. F.; Kaczmarek, M. E.; Ward, L. M.; Swanson, P. K; Lee, M.-C.; Marx, C. J. Development of an Optimized Medium, Strain and High-Throughput Culturing Methods for *Methylobacterium extorquens*. *PLoS One* **2012.** In preparation.
10. Delaney, N. F., Kaczmarek, M. E.; Marx, C. J. Curve Fitter: Open-Source Software for Minimizing Sources of Bias and Variance in Microbial Growth Curves. *PLoS One* **2012.** In preparation.
11. Standard ECMA-335: Common Language Infrastructure (CLI). *ECMA (European Association for Standardizing Information and Communication Systems)*, Geneva, Switzerland, **2002**.
12. Clarity's web page. http://code.google.com/p/osla/. Accessed June 2012.
13. Sparkes A; Clare, A. AutoLabDB: A Substantial Open Source Database Schema to Support a High-Throughput Automated Laboratory. *Bioinformatics* **2012**, *28*, 1390–1397.
14. Bär, H.; Hochstrasser, R.; Papenfulß, B. SiLA: Basic Standards for Rapid Integration in Laboratory Automation. *J. Lab. Autom*. **2012**, *17*(2), 86–95.

15. Overlord3, Real-Time Static Dynamic Scheduling Laboratory Automation Software. http://www.paa.co.uk/labauto/products/software/p-overlord3.asp. Accessed June 2012.
16. Gentsch, J. Flexible Laboratory Automation to Meet the Challenge of the ’90s. *Chemometrics and Intelligent Laboratory Systems* **1993**, *21*, 229–233.
17. Benn, N. D.; Liscouski, J. Discussion of Open-Source Methodologies in Laboratory Automation. *J. Lab. Autom*. **2009**, *14*, 82–89.

Supplemental Information for Chapter 4

Creating barcoded and fluorescently labeled strains

To create a genetic barcoding system to distinguish between otherwise identical strains, several plasmids derived from the pLW17 plasmid used to remove the cellulose operon were created so that, in addition to removing the *cel* locus to prevent biofilm formation, they would also introduce a unique sequence 44 basepairs in length. These sequences each contained an *Eco*RI recognition site at different locations. This was originally designed so that strains with different sequences could be distinguished amplifying genomic DNA by PCR and then performing a restriction digest (i.e., terminal restriction length polymorphism, or tRFLP). Sixteen unique sequences and derived plasmids were synthesized, leading to a total of 17 different plasmids (Table 4.1).

Each of these plasmids was then used to introduce a unique barcode sequence in to two strains of AM1 that expressed different fluorescent proteins. One strain, CM1175, expressed the mCherry protein while another, CM1179, expressed the Venus protein. These strains have been described elsewhere [1], and the proteins they express have distinct excitation and emission spectra so they can be readily distinguished. Together, this combination of 17 barcode types and 2 fluorescent protein types led to a collection of 17 x 2 = 34 uniquely identifiable strains.

To verify that these strains were selectively neutral relative to each other under the growth conditions in these experiments, 3 replicate fitness assays for each marked strain were performed that competed it against another. The presence of the barcode sequence in the strain was also verified by sequencing. Not all barcodes created initially appeared neutral, and so additional attempts were necessary to generate some of the desired strains. To begin this experiment, we used 19 of the 34 generated strains that were unambiguously verified in the first rounds of sequencing and fitness competitions.

Growth conditions, experimental layout and strains used

Populations of *M. extorquens* AM1 strains were evolved by batch culture in 48-well plates. Each population was founded using an equal ratio of strains that expressed either Venus or Cherry fluorescent proteins. To avoid contamination between different evolving populations, they were positioned in a checkerboard pattern on the 48-well plates so that every other well was only contained media that had not been inoculated (Table S4.1). To allow us to detect any contamination that might still occur, we founded the populations with pairs of Venus and mCherry expressing strains that had different genetic barcodes (Table S4.1).

Table S4.1 - Positioning of strains in 48-well plates at the start of the evolution experiment.

Eight different pairs of *M. extorquens* AM1 strains composed of one strain that expressed a Cherry fluorescent protein and another that expressed a Venus protein were used to found populations for the evolution experiment. The 8 different pairs are identified by roman numerals and their layout in a checkerboard pattern is shown in table A below. The barcodes used for each group are defined in the table B. To help monitor the populations through time and identify any problems, the corner wells were populated by either a Venus or a mCherry expressing strain (CM3120 or CM3140). This allowed calibration of relative fluorescence of each marker type on the plate reader, allowing us to detect any large shifts that occurred during the experiment in the frequency of the markers in each well.

A) Layout of 8 pairs of strains.

Row	Column 1	2	3	4	5	6	7	8
A	Control	IV		II		VIII		VI
B	I		VII		V		III	
C		V		III		I		VII
D	II		VIII		VI		IV	
E		VI		IV		II		VIII
F	III		I		VII		V	Control

Table S4.1 (Cont.)

B) Barcodes used for each of the strain combinations.

Group	Venus Strain Barcode	Cherry Strain Barcode
I	E-2	E-4
II	E-4	E-10
III	E-12	E-20
IV	E-10	E-14
V	E-8	E-12
VI	E-14	E-22
VII	E-16	E-28
VIII	E-20	E-10

The evolving populations were transferred every 48 hours. All populations were grown in 640 µL of MP medium per well. Populations were grown under two different demographic regimes created by altering the carbon substrate concentration and the dilution factor used when the populations were transferred. The large population environment was created by growing the cultures at a concentration of 20 mM methylamine·HCl and using a dilution factor of $\frac{1}{64}$ at each transfer. In contrast, the small populations were only grown at 3.33 mM methylamine·HCl and diluted by a ratio of $\frac{1}{4096}$ at each transfer.

For each population size, 96 separate populations were evolved. This led to a total of 182 populations distributed across eight 48-well plates. To found these populations with independent lineages and minimize the risk of pre-existing beneficial mutations, that were already present in the source culture, simply increasing in frequency during the evolution experiment, populations were started from master plates formed by picking single colonies. The source strains to be evolved were plated and individual colonies were picked and grown up separately in 48-well plates. These individual colonies were then mixed in a 50:50 ratio to create the starting populations. We created 96 such initial mixes of independent strains, and used each to start a large and small population (so that each selected colony was used to found both a large and small population). We used each single colony to found 2 evolving populations so that we could check for unusual convergence between populations founded by the same plated colony as an indication that mutations which occurred prior to the start of the evolution experiment were consequential.

Determining absolute population sizes

In this experiment, the ratio of the population sizes used is determined exactly by the substrate concentrations and dilution ratios in the experimental protocol. Since the DFE is a distribution of relative probabilities, this means that its inference is not affected by the actual population sizes used in this experiment, but only their ratios. However, in order to obtain information about the absolute beneficial mutation rate at which any mutation appears and not just their relative probabilities conditioned on a mutation appearing, we wanted to infer the actual population sizes present at the start and end of every transfer.

To do this, we counted the number of colony forming units (CFUs) that appeared on dilution plates. We grew a culture of *M. extorquens* in a 48 well plate on 17 mM methylamine. Three 100 µl samples were taken from the total of 640 µl in this well and each was serially diluted by a factor of 1e-5 before being plated at 10, 20 and 40 µl. CFUs for each of the resulting 3 x 3 = 9 plates were then counted and were all between 35 and 138 per plate. We assumed that the counts on each of these plates was Poisson distributed with a common mean after accounting for the different dilutions (a model which we could not reject, p-value = 0.46) and therefore estimated the total number of cells in the well as 2.15×10^8 (95% C.I. 1.99-2.31 x 10^8). By scaling this estimate at a 17 mM concentration to the 20 and 3.33 mM concentrations used in our experiments, leads to estimated final population sizes after growth of 2.53×10^8 and 4.22×10^7 for the large and small populations respectively.

Growth rate assays

Growth rates of evolved isolates as well as the population endpoints were determined using the instrumentation and models described previously. Cells were grown on 20 mM methylamine

after being inoculated from a 1/1000th dilution from stationary phase. The exponential model was fit over a range of OD values going from 0.02 to 0.2 (roughly 75% of the maximum value).

Fitness assays

Competition experiments

Competitive fitness differences between evolved isolates and ancestral strains were determined using the classic method of observing how the relative ratio of the two changes as they are co-cultured under the conditions of the evolution experiment. For each of the evolved isolates, three replicate competition experiments were performed. Each isolate was competed against an ancestral strain that expressed the alternative fluorescent protein (either Venus or mCherry). The ancestral strain and evolved isolate were first each acclimated for one growth cycle under the conditions of the evolution experiment for one transfer. Each type was then mixed in a 7:3 evolved to ancestral ratio and passaged to a new culture at the appropriate dilution. The mixed culture was then transferred twice more, creating three time-points with which to measure fitness.

Fitness was determined by estimating the frequency of each type after a cycle of growth using a flow-cytometer as described in the next section. Given these frequencies, the fitness difference estimated between any two time-points was calculated using a standard equation [2]. Denoting D as the relative increase in the population size between two time-points and F_0 and F_1 as the as frequency of an evolved isolate at the start and end of an experiment respectively we calculated fitness as:

$$W = \frac{\log\left(\frac{F_1 D}{F_0}\right)}{\log\left(\frac{(1-F_1)D}{1-F_0}\right)}$$

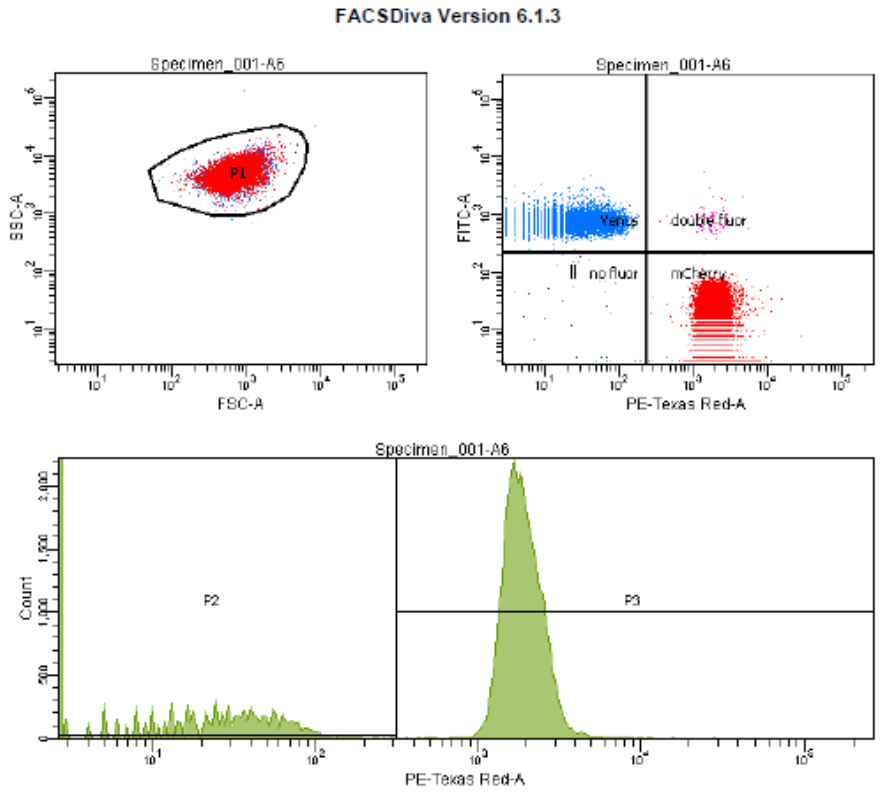
The three time-points allowed for $\binom{3}{2} = 3$ fitness values to be calculated between all time-points, denoted as W_{12}, W_{23} and W_{13}. We only reported the average of the values determined from the last and first time-points in the results, but used discrepancies between the first and second transfer intervals to add a as well as between replicates as metrics for quality control.

Counting fluorescent cell types by flow-cytometry

A BD LSRII flow cytometer (BD Biosciences, San Jose, CA, USA) was used to measure the ratio of fluorescently labeled cells in samples from fitness competitions with the initial data analysis being performed with the BD FACSDiva Software version 6.1.3. Samples were counted for a total of 50,000 events and at each event the forward and side scatter, as well as two fluorescent readings, were recorded. At the first stage of analysis the side scatter and forward scatter of all events was analyzed to exclude all events (typically <1%) that did not have values typical of a measured cell and that were assumed to represent either instrument "noise" or small particles. The remaining events were then categorized based on their fluorescent signal into one of four distinct groups, those with no fluorescent signal (NF), those with a Cherry signal (C), those with a Venus signal (V) and those with both a Venus and a Cherry signal (DF) (Fig S4.1).

An analysis of these four groups was then used to generate the ratio of Venus to Cherry cells in any sample. Events that fell into the NF group were assumed to represent either small particles in the media or instrument noise as the number of events that appeared in this group was not significantly different when an equivalent volume of blank media was analyzed, indicating that it is not a result of insufficient fluorescent signal. A small number of events (<1%) would typically

fall into the DF group (showing both Venus and Cherry fluorescence) and were attributed to the simultaneous passage of a Cherry and a Venus labeled cell being detected as a single event instead of two separate events. This interpretation is justified because no events are recorded in the DF group when pure cultures of Venus or Cherry labeled cells were analyzed, and the number of events in the group increased at smaller sample dilutions that resulted in a higher rate of events per second. The presence of two cells being counted as two events, as indicated by events in the DF group, does introduce a slight bias into the results and so we corrected for this bias using the likelihood procedure described next. Although larger dilutions could also ameliorate this bias, in practice diluting the sample enough so that no DF events would occur and the same total number of events could be counted would require using more liquid than the flow cytometer's sample holder could accommodate and would also result in prohibitively long analysis times per sample.



Experiment Name: Competition Assays PKS
Plate Name: I3_T1C_2
Specimen Name: Specimen_001
Well Name: A6
Record Date: Sep 26, 2012 10:27:07 AM
BOP: pswanson
GUID: 2397175a-f28a-4f3f-afee-f09eea44e560

Population	Parent Name	#Events	%Parent	FSC-A Mean	SSC-A Mean
All Events	####	50,000	####	742	4,801
P1	All Events	49,991	100.0	742	4,798
P2	P1	19,631	39.3	737	4,788
P3	P1	30,360	60.7	745	4,805
Venus	P1	19,563	39.1	737	4,786
double fluor	P1	103	0.2	1,495	9,530
no fluor	P1	62	0.1	832	5,281
mCherry	P1	30,263	60.5	742	4,789

Figure. S4.1 -Worksheet showing flow-cytometry gating method. The worksheet below is the typical example of the analysis files produced by the FACSDiva Software used to measure ratios of fluorescent cell types in the fitness competitions. The plot in the top left shows the side scatter and forward scatter range required to be defined as a countable cell. Also looms range are then assigned to one of four quadrants based on emission of either venus or cherry signal as shown in the upper right

As such, we instead modeled the counts of events in each of the three fluorescent groups as the sum of events due to single and double cell counts and corrected the bias due to double events using the equations below. We assumed that all events were due to only one or two cells moving in front of the detector because if one assumes that cells hit the detectors as a Poisson process with constant rate λ and events result in a DF categorization if one or more cells hits the detector within a window length of time t after an initial event, then if approximately 1% of all events result in DF categorizations, <0.6% of these events will be due to more than 2 cells, indicating that this is a very safe approximation. In fact, in general since the DF events are rare, accounting for the presences of the DF group in calculating the ratio only changes the frequencies by less than 1 percent compared to a simple calculation that only used the ratio of the two fluorescent groups. Using this assumption and the following notation the events in each group are then defined as:

$$p_D = Proportion\ of\ counted\ events\ measuring\ two\ cells$$

$$p_S = 1 - p_D = Proportion\ of\ counted\ events\ measuring\ one\ cell$$

$$p_C = Proportion\ of\ Cherry\ labeled\ cells\ in\ a\ sample$$

$$p_V = 1 - p_C = Propotion\ of\ Venus\ labeled\ cells\ in\ a\ sample$$

$$N_C, N_V, N_{DF} = Number\ of\ Cherry, Venus\ and\ Double\ Fluorescent\ events\ counted$$

$$p_{N_C}, p_{N_V}, p_{N_{DF}} = Probability\ of\ a\ Cherry, Venus\ or\ Double\ Fluorescent\ event$$

$$p_{N_C} = p_C^2 p_D + p_S p_C$$

$$p_{N_V} = p_V^2 p_D + p_V p_S$$

$$p_{DF} = 2 p_V p_C p_D$$

The likelihood of the data conditioned on the total number of counts is thus a multinomial form and proportional to:

$$l(N_C, N_V, N_{DF}|p_C, p_S) \sim N_C \log[\, p_C^2 p_D + p_S p_C] + N_V \log[p_V^2 p_D + p_V p_S] + N_{DF} \log[\, 2 p_V p_C p_D]$$

Giving the maximum likelihood estimates for p_C and p_V are then obtained from the data as:

$$\hat{p}_C = \frac{2N_C + N_{DF}}{2(N_C + N_V + N_{DF})}$$

$$\hat{p}_V = \frac{2N_V + N_{DF}}{2(N_C + N_V + N_{DF})}$$

And these estimates were used in the corresponding fitness equations.

Estimates of Venus and Cherry cell-type frequencies in populations at the end

Endpoint frequencies of the Venus and Cherry marker were obtained for the entire populations using an identical counting mechanism as the fitness assays.

Population Genetic Model and Statistical Inference

We used the fitness of isolates obtained at the end of the experiment from each replicate population to fit a discretized version of the DBFE in a Bayesian framework using Gibbs sampling. The guiding principle of the algorithm described below is that it would be a simpler problem to estimate the DBFE and the beneficial mutation rate if one knew all the beneficial mutations that occurred throughout the experiment and what their fitness effects were. We therefore treated these unobserved mutations as latent variables that are integrated over by Gibbs sampling. A full description of the algorithm is given below.

1. Discretize the DBFE into L bins or fitness classes, including a point at 0 to represent neutrality, and equally spaced values from W_{Min} to W_{Max} with the fitness of each class being the midpoint of the interval, $\delta_i, i = 1 \ldots L$. The isolate fitness values, the observed data, are assumed to come from the fitness class whose value is closest to theirs, and are counted as observations from that category. If the fitness of the isolate is "neutral," then it is counted as coming from the "0" class.

2. Draw the vector of relative probabilities for each class in the DBFE with a positive selection coefficient, $\boldsymbol{p} = \{\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3, \ldots \mathrm{p}_\mathrm{L}\}$ from a Dirichlet prior distribution, also draw a beneficial mutation rate, μ_b, from a Gamma prior distribution.

3. For each population and isolate combination, sample a new set of beneficial mutations that occurred during the evolution experiment based on the current parameter values. This Gibbs step is implemented by rejection sampling where the simulation is discarded if it does not yield the observed fitness for an isolate, and accepted as a sample if it does. Further details for this simulation step are given in the next section.

4. Conditioned on the missing data of beneficial mutations and their selective effects for all replicate populations imputed in (3), sample a new beneficial mutation rate from the conditional posterior distribution, μ_b, which will be a draw from a $Gamma(\alpha + \sum t_i\,, \beta + M)$. This is the traditional conjugate posterior distribution for a Poisson process rate parameter with a $Gamma(\alpha, \beta)$ prior. The t_i represents the total number of doublings that occurred in a population during the evolution experiment, and M is the total number of mutations that occurred in all the populations.

5. Conditioned on the missing data in (3) sample a new vector or probabilities, $\boldsymbol{p}$, for the fitness classes in the discretized DBFE from a $Dirchlet(\alpha + M_1, \alpha + M_2, \ldots, \alpha + M_L)$, where M_i is the total number of mutations that appeared from that class. This distribution is the conjugate posterior for a multinomial with a $Dirichlet(\alpha, \alpha, \ldots, \alpha)$ prior distribution.

6. Return to (3) and repeat until enough samples are obtained to adequately estimate the posterior distribution of all model parameters.

Simulating evolving populations

For step (3) in the inference algorithm, it is necessary to simulate the evolution of each of the replicate populations. We modeled the growth of cultures in microtiter plates as a deterministic interval of continuous growth, during batch culture, followed by random discrete sampling during the transfer step. The deterministic dynamics are modeled as follows. The population starts at time 0 with all members having equal fitness. The ancestral type grows exponentially with rate $r_0 = \log(2)$, so that time is rescaled to the general time of the ancestor. A new mutation which appears from fitness class i grows per unit time with a rate $r_i = r_0 + \delta_i r_0$. We model the change in frequency of each of the different fitness classes as a constantly growing population, where the frequency of a fitness group with at a particular time, $f_i(t)$, is conditioned on the starting population size for each class, A_i:

$$f_i(t) = \frac{A_i e^{r_i t}}{\sum_{j=0}^{L} A_j e^{r_j t}}$$

In the simulation, the $A_i$values are set at the start of each transfer, and at the start of the simulation, only the A_0 class has any members and its size is set to the population size after the transfer, N_0, determined by the experimental protocol. Some additional modifications in the

simulation are made to account for the complication that time in the model is scaled to the ancestral generation time. Namely, as the population increases its fitness it exhausts all the substrate available during batch culture in less time than it did on the original timescale, shortening the period of time between transfers. To approximately compensate for this effect, the total time in between transfers is calculated by assuming that the entire population grows at the mean population growth rate that exists at the start of the transfer, $\bar{r}$, assuming it always grows to the same final population size, N_f, before being transferred. To model the transfer itself, the number of individuals after the transfer from each fitness class is sampled from a Poisson distribution whose mean is equal to the expected number from that class, which is its current population size multiplied by the dilution factor.

Against this background of different fitness classes changing their frequencies deterministically during growth, we stochastically add beneficial mutations. The stochastic appearance of beneficial mutations is separately simulated for each fitness class during batch culture. To perform this simulation, for each fitness class, we first define a time-scale on which mutations are introduced at a constant rate. Mutations are typically thought to occur during cell-division and to reflect this the rate that they appear should increase with the growth-rate of the organism. A time-scale that gives equivalent probability of a mutation occurring then, is one where the probability of a mutation is uniformly distributed amongst all equally spaced intervals on that scale. Consider the following transformation which converts actual time, t, to a new scale, g, given by $g = e^{tr_i}$. Then, $t = \frac{log(g)}{r_i}$, and as the population size of an exponentially growing population at any time is given by $N(t) = N_0 e^{r_i t}$, then for any interval, $[g, g + \Delta t]$ the population will have increased by $N_0 e^{r_i(\log(g+\Delta g))} - N_0 e^{r_i \log(g)} = N_0 \Delta g$. That is, there will be the same amount of growth in all intervals of Δg, satisfying the requirement.

With this new time-scale, the appearance of beneficial mutations can be simply simulated by a three-step process. First, the total number of mutations that appear in that fitness class during growth is simulated as a Poisson random variable whose mean is equal to the beneficial mutation rate times the amount of growth in between transfers t_k. The mean for each group, λ_j is thus $\lambda_j = \mu_b\left(N_f - N_0\right) = \mu_b(N_{j0}e^{r_j t_k} - N_{0j})$. In the second step, each of these mutations is given an appearance time by sampling them uniformly from the modified time-scale and then converting these times back to the original time-scale. Finally, the selective effect of each mutation is assigned by drawing it from the multinomial distribution of relative probabilities for the different fitness classes, $\boldsymbol{p}$. To model the redundant effect of different beneficial mutations that seemed to be indicated by the stationary phase behavior, a lineage with two or more beneficial mutations had its fitness value set to the maximum individual effect of any mutation.

Each beneficial mutation that appears grows deterministically until the next transfer. A mutation appearing at time t_{new} therefore has $t_k - t_{new}$ of time to grow from an initial size of one. So we add $e^{r_i(t_k - t_{new})}$ to the population of the fitness class this mutated lineage belongs to before the sampling step. To compensate for the slight increase in the population size due to the growth of lineages representing new mutations, the final frequency of the fitness class the mutant appeared in is reduced by an amount equal to the total growth of the advantageous mutant. One the populations have evolved for the number of transfers set by the experimental protocol, an isolate is sampled from the population according to the different frequencies of each fitness class. If the isolate matches the observed fitness, the simulation is considered a valid sample, if not, it is discarded and the entire simulation is repeated. Source code to perform these simulations is available from the authors.

Table S4.2 - Substrate concentrations, dilution ratios and effective population sizes for the different populations. The population sizes used for this evolution experiment. The effective population size is calculated as explained next.

Population Size	Population size before transfer	Dilution ratio	Size After Transfer	Effective Population Size
Large	2.53×10^8	1/64	3.95×10^6	1.64×10^7
Small	4.22×10^7	1/4096	1.03×10^4	8.56×10^4

Effective population sizes and why we did not use them.

Past experimental work investigating fitness effects has been done using populations that evolve under a regime similar to the type of batch transfer culturing used here, large expansions followed by periodic bottlenecks. In contrast, most of these experiments are analyzed after assuming the evolving population was always a constant size, and this size is referred to as an effective population size. The link that has allowed for this is the assumption that the dynamics of these fluctuating populations can be modeled by an equivalent population of constant size.

The key question when determining if a population that is constantly changing is equivalent to one that is constant in size is what one means by equivalent. Typically, a single quantitative aspect of the population, such as the variance in the change in allele frequencies each generation, is found to be identical in both populations to justify the simplification to an "effective" population. For experiments that have evolved populations to look for beneficial alleles, the typical parameter of interest is usually the probability that a beneficial mutation escapes stochastic loss. These populations were therefore approximated by a constant population where this probability is equal.

A formula exists to determine, given the pattern in which a population dynamically changes, the size of a constant population where mutations are as equally likely to escape stochastic loss [3]. This is a fantastic approximation for small fitness effect sizes, and has the added benefit of agreeing well with the effective population size as defined with respect to the variance in allele frequency changes. However, although this approximation is accurate for a regime where the actual population does not greatly change or where most mutational effects are small, it can break down if these conditions are not met and it does systematically bias the DBFE. Further, for the inference of the distribution of fitness effects, not only the odds that a mutation escapes

drift, but also the time for it to increase in frequency are of interest. Mutations that escape drift in most constant sized population models that are stochastically sampled every generation often rise in frequency faster than could be deterministically expected, because by escaping drift they have gotten “lucky” and so usually have had an atypically large number of progeny in the first several generations after appearing in the population. Evolutionary dynamics such as this, which can affect the inference model, are not always equivalent between two models that differ in whether the population size is constant or periodically expanding and contracting.

To simply avoid these complications, when performing inference we did not model the dynamics of the constantly changing populations used in this experiment with a constant population size. However, for interpretability, we do report the population sizes used in the experiments with a single value using the standard formulation for effective population sizes [3].

References

1. Lee MC, Chou HH, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of Methylobacterium to distinct growth substrates. Evolution 63: 2816-2830.

2. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332: 1190-1192.

3. Wahl LM, Gerrish PJ (2001) The probability that beneficial mutations are lost in populations with periodic bottlenecks. Evolution 55: 2606-2610.

Chapter 5 Back Matter

Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*

A study of evolutionary rates and processes following a host shift.

Reprinted from PLoS Genetics

Delaney, Nigel F., et al. "Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, *Mycoplasma gallisepticum*." *PLoS genetics* 8.2 (2012): e1002511.

Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*

Nigel F. Delaney[1], Susan Balenger[2], Camille Bonneaud[1¤a], Christopher J. Marx[1], Geoffrey E. Hill[2], Naola Ferguson-Noel[3], Peter Tsai[4], Allen Rodrigo[4¤b], Scott V. Edwards[1]*

1 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **2** Department of Biological Sciences, Auburn University, Auburn, Alabama, United States of America, **3** Poultry Diagnostic and Research Center, University of Georgia, Athens, Georgia, United States of America, **4** Bioinformatics Institute, University of Auckland, Auckland, New Zealand

Abstract

Measureable rates of genome evolution are well documented in human pathogens but are less well understood in bacterial pathogens in the wild, particularly during and after host switches. *Mycoplasma gallisepticum* (MG) is a pathogenic bacterium that has evolved predominantly in poultry and recently jumped to wild house finches (*Carpodacus mexicanus*), a common North American songbird. For the first time we characterize the genome and measure rates of genome evolution in House Finch isolates of MG, as well as in poultry outgroups. Using whole-genome sequences of 12 House Finch isolates across a 13-year serial sample and an additional four newly sequenced poultry strains, we estimate a nucleotide diversity in House Finch isolates of only ~2% of ancestral poultry strains and a nucleotide substitution rate of $0.8-1.2\times10^{-5}$ per site per year both in poultry and in House Finches, an exceptionally fast rate rivaling some of the highest estimates reported thus far for bacteria. We also found high diversity and complete turnover of CRISPR arrays in poultry MG strains prior to the switch to the House Finch host, but after the invasion of House Finches there is progressive loss of CRISPR repeat diversity, and recruitment of novel CRISPR repeats ceases. Recent (2007) House Finch MG strains retain only ~50% of the CRISPR repertoire founding (1994–95) strains and have lost the CRISPR–associated genes required for CRISPR function. Our results suggest that genome evolution in bacterial pathogens of wild birds can be extremely rapid and in this case is accompanied by apparent functional loss of CRISPRs.

Citation: Delaney NF, Balenger S, Bonneaud C, Marx CJ, Hill GE, et al. (2012) Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*. PLoS Genet 8(2): e1002511. doi:10.1371/journal.pgen.1002511

Editor: David S. Guttman, University of Toronto, Canada

Received August 13, 2011; **Accepted** December 9, 2011; **Published** February 9, 2012

Funding: This work was supported by National Science Foundation (http://www.nsf.gov/) grant DEB-0923088 to GEH and SVE, a NSF graduate fellowship to NFD, and a Milton Grant through Harvard University to SVE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sedwards@fas.harvard.edu

¤a Current address: Station d'Ecologie Expérimentale du CNRS a Moulis USR 2936 Moulis, Saint-Girons, France
¤b Current address: National Evolutionary Synthesis Center, Durham, North Carolina, United States of America

Introduction

Populations of animals are under constant threat from bacterial pathogens, which can be particularly destructive following a switch to a new host or the evolution of novel virulence mechanisms. Understanding the rate and process of evolutionary change in pathogens is thus important to assessing the risks of pandemics and developing means to predict and avoid such catastrophic events. In 1994, a strain of *Mycoplasma gallisepticum* (MG) was identified as the causative agent of an emerging epizootic in House Finches, a wild songbird inhabiting Eastern North America [1]. This bacterial pathogen frequently causes disease in commercial chicken and turkey flocks, but it had never been reported in House Finches or any songbird, leading to the suggestion that the epidemic began when MG expanded its host range from poultry to this phylogenetically distant songbird. MG prevalence reached 60% in some areas, and killed an estimated 225 million finches in the first three years after detection [2]. The early detection of the epizootic allowed research and citizen-science teams to track its rapid spread throughout eastern North America in exceptional detail, making it one of the best documented wildlife pathogen outbreaks [3–7].

Although previous genome-wide studies have clarified rates of measurable evolution in viral pathogens [8,9] and in bacterial populations evolving under laboratory conditions or as human pathogens [10–18], less is known about rates of genetic change in bacterial pathogens of non-mammalian vertebrates, particularly on short evolutionary time scales. Genome-wide and gene-specific estimates of point substitution in bacterial lineages measured over centuries [19] to millions of years [20] suggest maximum substitution rates on the order of 10^{-7} to 10^{-9} per site per year. Although recent work suggests the rate may be even faster for several bacterial species [12,14,19], the number of studies documenting whole-genome changes in bacteria during host switches is still small, particularly for wildlife pathogens [21,22]. Asx part of ongoing surveillance, field isolates of MG obtained from infected finches were sampled at multiple time points from the start of the epidemic in 1994 to 2007, providing a genetic time

Author Summary

Documenting the evolutionary changes occurring in pathogens when they switch hosts is important for understanding mechanisms of adaptation and rates of evolution. We took advantage of a novel host–pathogen system involving a bacterial pathogen (*Mycoplasma gallisepticum*, or MG) and a songbird host, the House Finch, to study genome-wide changes during a host-shift. Around 1994, biologists noticed that House Finches were contracting conjunctivitis and MG from poultry was discovered to be the cause. The resulting epizootic was one of the best documented for a wildlife species, partly as a result of thousands of citizen science observers. We sequenced the genomes of 12 House Finch MG strains sampled throughout the epizootic, from 1994–2007, as well as four additional putatively ancestral poultry MG strains. Using this serial sample, we estimate a remarkably high rate of substitution, consistent with past implications that mycoplasmas are among the fastest evolving bacteria. We also find that an array of likely phage-derived sequences known as CRISPRs has degraded and ceased to recruit new repeats in the House Finch MG strains, as compared to the poultry strains in which it is diverse and rapidly evolving. This suggests that phage dynamics might be important in the dynamics of MG infection.

series beginning immediately after the host switch, as well as an opportunity to directly measure the tempo and mode of evolution in a natural bacterial population whose genome is as yet uncharacterized.

To characterize patterns of genomic change during its host switch between distantly related avian species, we sequenced whole genomes of 12 House Finch MG isolates from this 13-year time series, with four samples each from the beginning (1994–1996), middle (2001) and recent (2007) periods (Table S1). In addition, to identify putative source strains as well to determine if differences between the House Finch MG strains and the ~1 Mb published reference R_{low} strain from chicken [23] were ancestral or derived, we sequenced four additional strains from chicken and turkey based on phylogenetic analysis of a smaller multistrain data set (Figure S1). Our sequence, SNP filtering and between-platform cross-validation protocols yielded a high quality 756,552 bp alignment encompassing 612 genes (Tables S2, S3, S4, Text S1, Figure S2), and allowed us to monitor point substitutions, genomic indels, IS element insertions, and other changes across the entire genome (Figure 1), including the entire array of clustered regularly interspaced short palindromic repeats (CRISPR) of all 17 strains (finch and poultry isolates).

Results

Phylogenomic diversity of House Finch and poultry MG

All House Finch MG samples were collected in the southeastern U.S. (Table S1), with an emphasis on the well studied population in Alabama [24,25]. The population structure of Eastern House Finches before the epizootic was virtually panmictic [26], suggesting that there is likely to be little geographic structuring of MG in the east, a hypothesis that could be tested with additional data. The 12 House Finch strains from the three time periods spanned the known temporal and phylogenetic diversity of this lineage, and included strains that have been used to study host response to pathogen infection in House Finches [27]. To determine genetic diversity and phylogenetic identity of putative source populations of the House Finch MG strains, and to aid in sampling chicken and turkey strains for sequencing, we first analyzed a previously published data set [28]. Phylogenetic analysis of 1,363 bp obtained from four genomic regions for a large sample ($n = 82$) of MG strains suggests that turkeys rather than chickens were the source of House Finch MG and that the MG lineage colonizing House Finches first passed multiple times among chickens and turkeys (Figure S2). Although this analysis suggests frequent host switches between chickens and turkeys, which diverged 28–40 MYA [29,30], it also suggests a single switch to the House Finch, a songbird species diverged from chickens by ~80 MYA [31].

The whole genome alignment contained strong signals of a founder event as a result of colonization of House Finches. The total nucleotide diversity (π) in the House Finch strains for the four-gene region was only 3.1% of the diversity in circulating poultry strains prior to the epizootic, and only 2.3% of the poultry diversity when considering the entire House Finch MG genome [28] (Figure 2 and Table S5). In agreement with the four-gene analysis, our whole genome sequencing showed that the four sequenced poultry isolates were much more genetically diverse than the 12 House Finch isolates, possessing a total of 13,175 SNPs as compared to only 412 SNPs among the House Finch isolates (Table S2). The House Finch MG diversity corresponds to $\pi = 0.00014$, or roughly 1 SNP every 1,800 bp. Consistent with purifying selection acting over the longer time period encompassing the divergence of House Finch and poultry MG strains (as opposed to acting after the host-switch among House Finch strains alone), there was a stronger bias against non-synonymous substitutions among the more diverged poultry strains than among the recently diverged House Finch MG strains (Table S6). Across the entire genome, only 147 (35%) of the SNPs among the House Finch isolates were phylogenetically informative; the majority (265 or 64%) appeared as singletons.

To further quantify House Finch MG demography, we used a statistical model, the Bayesian skyline plot implemented with BEAST, that utilizes information on dates of sampling to estimate changes in genetic diversity through time [32,33] (Text S2). The analysis is broadly consistent with field observations suggesting a mid-1990s origin followed by rapid population expansion, though it estimates that the House Finch MG lineages coalesced roughly in 1988, several years prior to the observation of sick birds in the field (estimated MRCA of the House Finch MG strains is 19.2 years prior to 2007 [95% HPD 16.9 – 21.7]; Figure 2d). Discrepancies between coalescence times and observed outbreaks in host populations have been observed for other pathogens, and could possibly be due to selective or demographic effects, or in our case low sample size [12]. Phylogenetic analysis suggests substantial turnover in the standing SNP variation between sampling intervals, with strong clustering of the 2007 strains, which are distinguished from other House Finch strains by 85 diagnostic SNPs (Figure 3). We found that one of the sequenced turkey strains, TK_2001, was highly similar in sequence to the House Finch strains and shares a number of genomic deletions and transposon insertions as well as duplications and losses of CRISPR spacers (see below) with the House Finch MG strains. This turkey strain may represent a poultry lineage close to the source lineage for House Finch MG (Figure 3).

In addition to SNPs in House Finch MG we found five large genomic deletions that occurred by 2007 and amounted to ~42,245 bp and encompassing 34 genes relative to the chicken R_{low} strain (Figure 1 and Figure 3, Table S7). Three of these deletions are phylogenetically informative among the 17 MG strains (Table S7), but their conflicting phylogenetic distribution underscores the presence of recombination (see next section). Two deletions

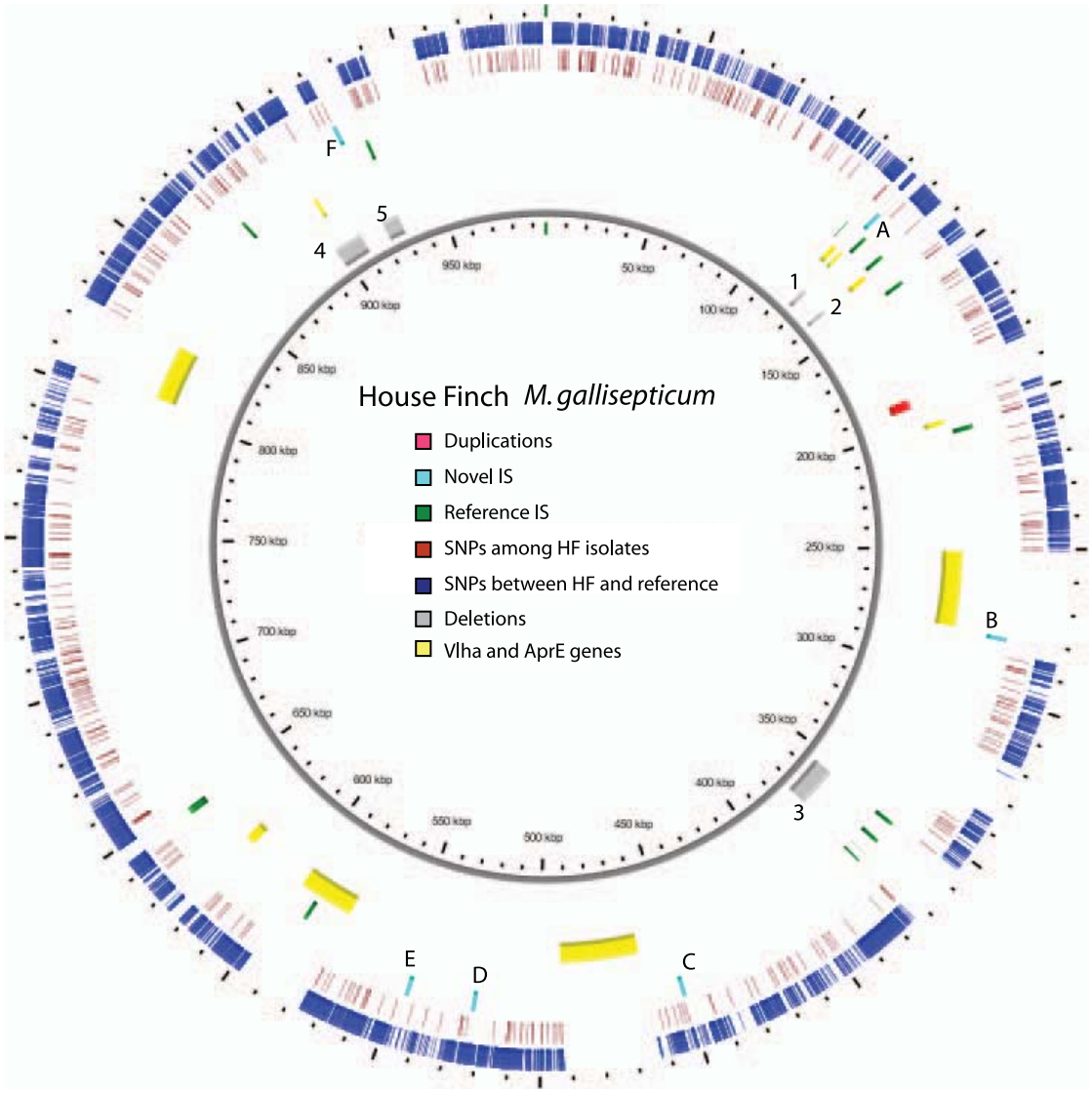


Figure 1. Overview of the genome of the House Finch strain of *Mycoplasma gallisepticum* summarizing variation among 12 House Finch MG isolates and comparing these to a poultry reference (0.99 Mb). Blue ticks indicate SNPs fixed within the House Finch isolates and differing from the chicken MG reference. Red ticks indicate polymorphisms among the House Finch isolates. Yellow regions are unassembled repetitive regions including VlhA and AprE genes. Grey regions indicate 4.8% of the aligned genome that is deleted in the House Finch isolates; numbers correspond to deletions detailed in Table S12. Green and light blue ticks indicate IS elements (family IS1634) in the reference genome and novel sites in the House Finch strains, respectively; letters next to novel sites correspond to insertions detailed in Table S9.
doi:10.1371/journal.pgen.1002511.g001

totaling 9,275 bp were shared among all strains except the reference. In addition, we detected six novel IS element insertions in the House Finch MG lineage (Text S3, Table S8) and three of the genomic deletions were likely mediated by illegitimate recombination between flanking IS elements (Table S7). In addition to the 34 genes deleted as part of genomic deletions, we found evidence for pseudogenization of 19 genes relative to the chicken MG reference (Text S3, Table S9). Two genes appear to have been disrupted by transposon insertions and 17 genes were pseudogenized by frameshift or nonsense mutations (Table S9). The substantial gene losses we detected, a total of 52 genes (~8.6%) fixed in the House Finch MG lineage, presumably as a result of the bottleneck during host switch. By contrast, we failed to find a single novel gene in House Finch MG that was not also found in the poultry MG strains (Text S5). Comparative analysis with other *Mycoplasma* genomes showed that 15% of these lost genes also lacked a homologue in the other genomes surveyed whereas 13% had a homologue in every genome (Table S9).

Recombination and lateral gene flow

Despite the small amount of genetic variation segregating among our House Finch *Mycoplasma* samples (only 412 SNPs), it is not possible to construct a phylogenetic tree for these strains that is free of homoplasies. Although the four 2007 strains and all 2001

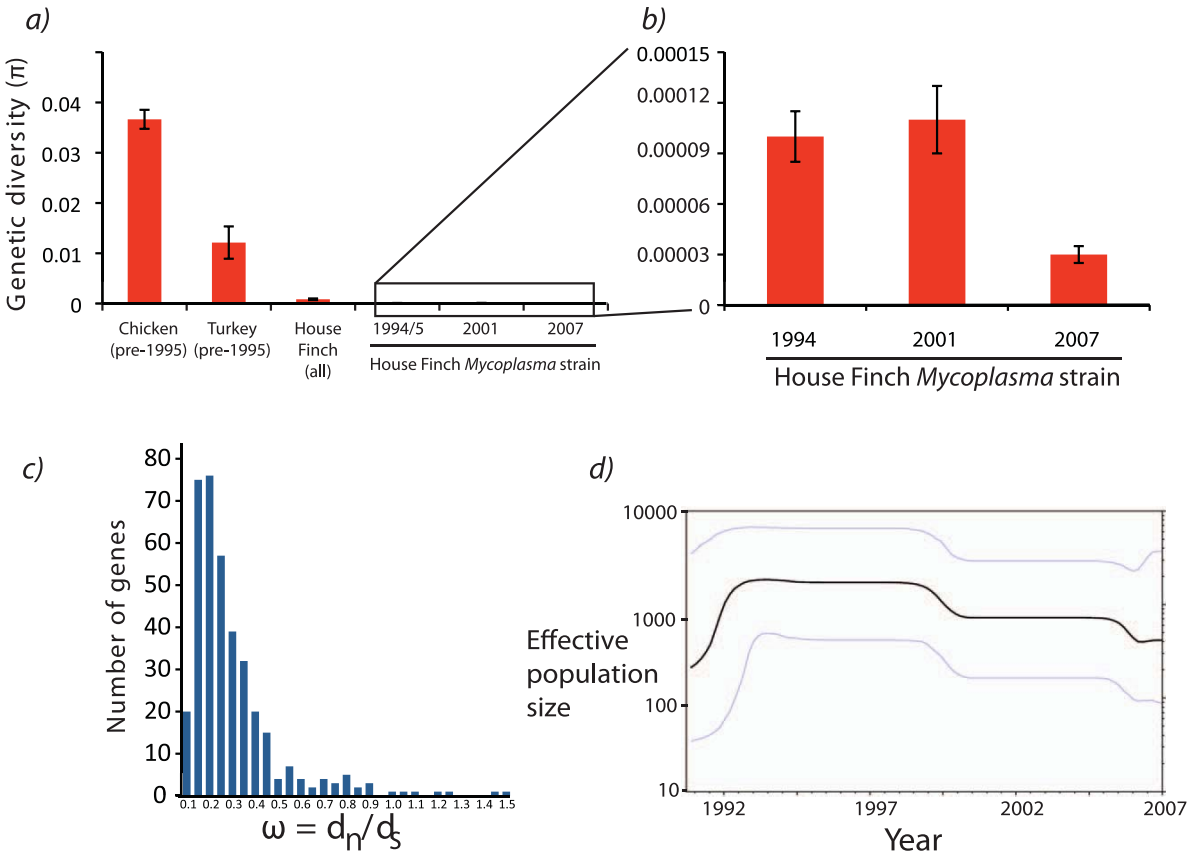


Figure 2. Patterns of polymorphism among *Mycoplasma gallisepticum* isolates collected from House Finches. a) Comparison of nucleotide diversity between historical chicken MG strains and serially sampled House Finch MG isolates for a 1.3 kb region [28]. b) Expansion of House Finch nucleotide diversity measured across the whole-genome alignment (approximately 738 kb when considering only the 12 House Finch isolates). c) Patterns of synonymous and nonsynonymous substitution for all MG isolates sequenced in this study as well as the reference. The values in this histogram reflect estimates of $\omega = d_n/d_s$ across a tree including all House Finch isolates and the poultry R_{low} reference. For a full list of patterns of substitution for each gene, see Data S1 (Estimates of omega.xls). d) Bayesian skyline plot estimated from the alignment of 12 of house finch Mycoplasma strains. Although the upper and lower 95% confidence limits (gray lines) on the skyline plot are substantial, the overall trend (black line) is indicative of population growth approximately 17 years before 2007, or 1990, placing the spread of MG somewhat earlier than the first field observations in 1994. Note that time is reversed so that time proceeds from left (past) to right (most recent time of sampling).
doi:10.1371/journal.pgen.1002511.g002

strains except AL_2001_17 clearly formed well defined clades based on 85 and 28 SNPs, respectively, establishing the phylogenetic relationships for the other 5 House Finch MG strains exclusively via SNPs was not possible (Text S6, Figure 3). Although a total of 16 SNPs were phylogenetically informative for the placement of these five strains, the largest cluster of SNPs that were phylogenetically consistent was seven, and overall, 13 different trees were supported by at least 3 SNPs each. Similarly, substantial homoplasy was found among the four newly sequenced poultry strains and the R_{low} reference. Although 6,152 SNPs were parsimony informative for these five strains, the unrooted tree with the best support was in conflict with 4,619 (75%) of these SNPs. These patterns are expected if sites are being shuffled by recombination or horizontal gene transfer (HGT) among isolates, and analysis of the entire data set found strong support for this (Text S4, Figures S3, S4, S5). Using the pairwise homoplasy index test [34] revealed a statistically significant signal of recombination ($p<10^{-9}$). This signal comes predominantly from the four newly sequenced poultry strains because there is not enough genetic variation to make this test significant when only the House Finch strains are considered. However if we apply to the House Finch MG strains the homoplasy test by Maynard-Smith and Smith [35], which is found to perform well in situations of low nucleotide diversity [36], we again obtain a significant signal for recombination ($p<10^{-6}$). We conclude that, despite a significant signal for recombination in both the poultry and House Finch strains, the House Finch MG cluster as a whole is a distinct and easily identifiable phylogenetic lineage with a long branch separating it from the poultry strains (Figure 3).

Substitution rate and robustness to model assumptions

Coalescent analysis [32] of the 12 House Finch isolates sampled at different dates suggested an extraordinary point substitution rate of 1.02×10^{-5} substitutions per site per year (95% HPD $7.95\times10-6$ to $1.23\times 10-5$ (Text S2), consistent with earlier suggestions that *Mycoplasma* may be among the fastest evolving bacteria [37]. This rate of point substitution is not restricted to House Finch MG strains but was also found in the poultry strains when analyzed separately (Text S2), suggesting that rapid evolution was characteristic of MG prior to the House Finch

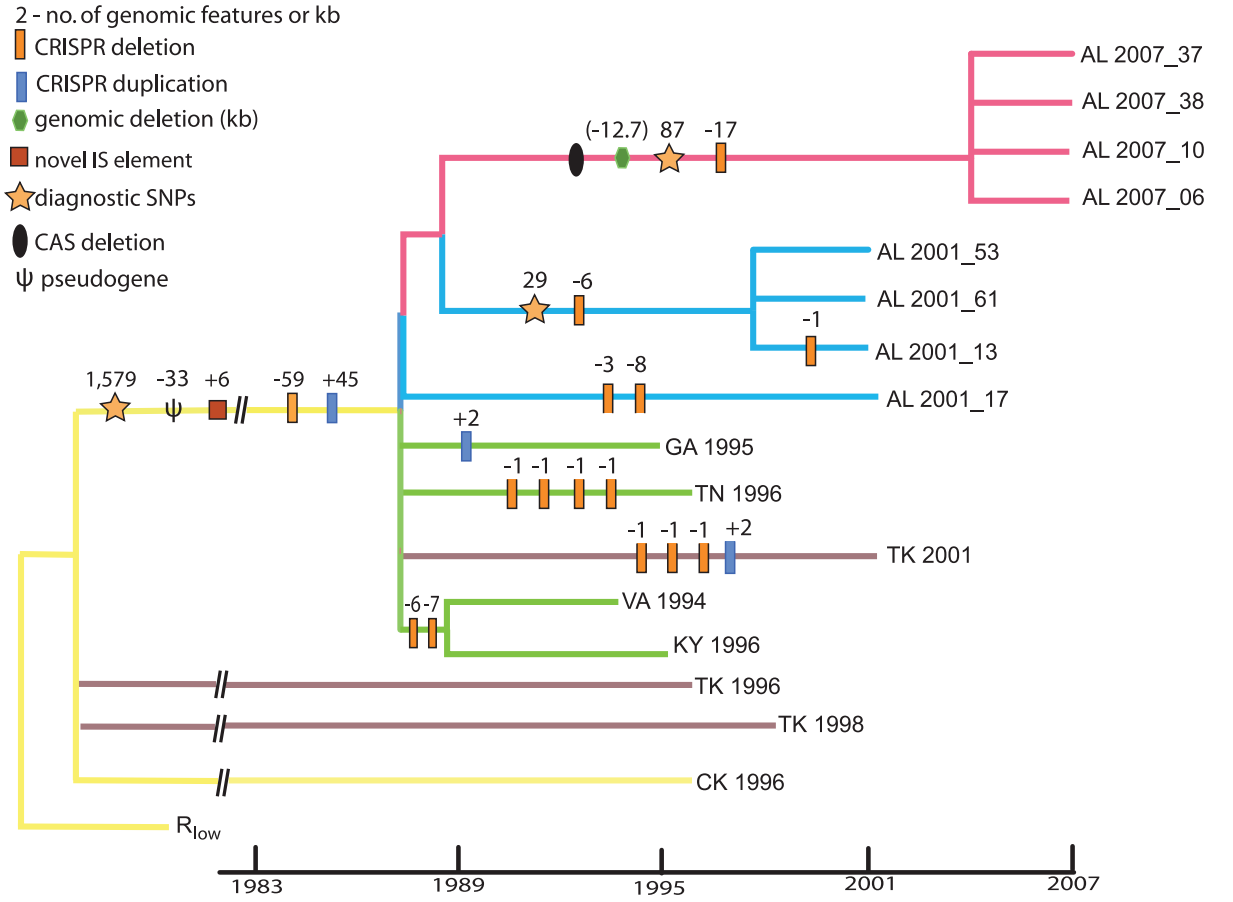


Figure 3. Phylogeny of *Mycoplasma gallisepticum* isolates collected at time points 1994–2007 following a host shift from poultry to House Finches. The basic topology and branch lengths of the tree come from the output for the BEAST analysis made while estimating evolutionary rates. From this tree we collapsed branches with less than 0.6 posterior probability or if there were no phylogenetically informative SNPs supporting that branch. Several strains are shown as polytomies because their genomic histories are shaped by recombination. Within the House Finch MG clade, branch lengths are proportional to time. Major genomic events are indicated on appropriate branches. The numbers of diagnostic SNPs indicated on various branches are minima. The numbers of CRISPR changes shown are only those that can be constructed with reasonable support (Figure 5); one possible reconstruction is presented.
doi:10.1371/journal.pgen.1002511.g003

epizootic. We estimated a similar substitution rate when considering only the four-gene multistrain alignment use to identify poultry strains for sequencing (Text S2). We verified that our estimate of substitution rate is robust to different protocols for SNP identification, statistical models and data sets (Figure 4; Text S7). Altogether we estimated the substitution rate within a coalescent framework on 34 combinations of SNP calling and model assumptions and found consistent estimates throughout (Text S1, Figure 4, Figure S6). In addition, we achieved a similar estimate using a Poisson regression approach as well as a root-to-tip regression (Text S7 and Figure 4).

A possible mutator strains in House Finch MG

In addition to a high estimated substitution rate in MG, we found a mutation in the gene-encoding *UvrB* that could elevate this rate yet further. *UvrB* is an essential part of the nucleotide excision repair system, which has been posited to be the most important pathway for maintaining genomic integrity in *Mycoplasma* [38]. The mutation truncates the *UvrB* protein by three amino acids (Table S10) and raises the possibility of the origin of a mutator strain in House Finch MG [39] as the C-terminal of this protein is essential for its function [40]. Consistent with this idea, we found 14 instances of adjacent SNPs among the 12 House Finch isolates, a notable excess in an alignment with only 412 variable sites (Table S11). Moreover, 12 of these 14 are CC→TT double substitutions, which are normally repaired by the UVR system (Table S10). For 13 of the 14 doublets, both sites are inferred to have mutated on the same branch of the tree, suggesting single mutational events, and the proportion of doublet mutations involving the same base was drastically higher (92.8%) in lineages with the *UvrB* mutation as compared to those without ($p<0.0001$; Table S10). Nonetheless, these doublet mutations are not required to achieve the high rate of substitution that we measured. They account for less than 7% of the segregating variation and removal of these doublet sites does not affect the high estimated substitution rate. The *UvrB* mutation is found in all of our House Finch MG strains as well as the turkey strain TK_2001, but not in the ancestral chicken strains or the reference chicken strain. Thus, the mutation appears to have arisen on the lineage leading to the House Finch.

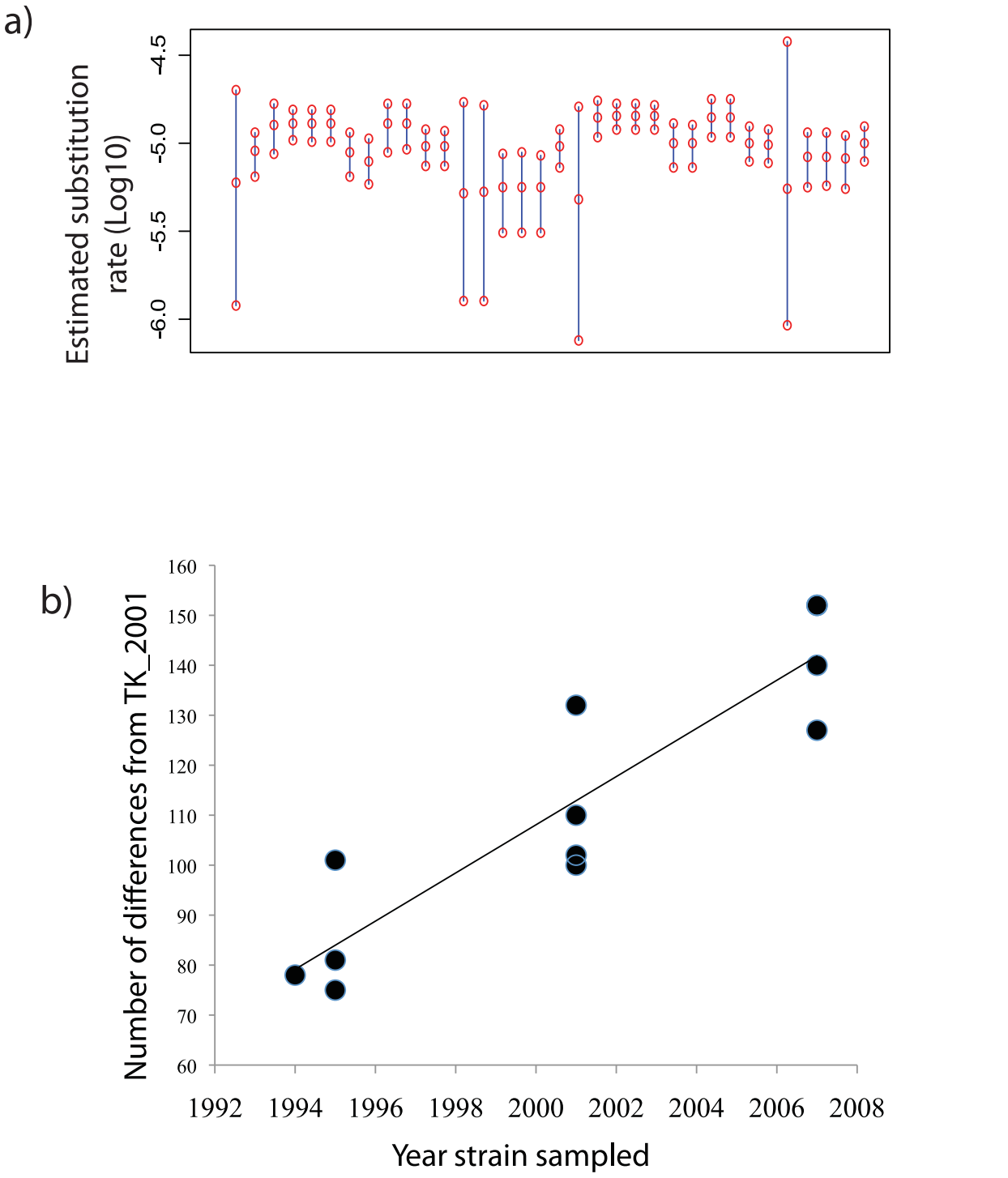


Figure 4. 95% highest posterior density intervals on the estimated substitution rate. A) for House Finch *Mycoplasma* strains derived from 34 analyses using the different data and model combinations described in Text S2. The middle circle of each bar is the estimated mean; top and bottom circles are the upper and lower 95% bounds of each highest posterior density (HPDs). b) Root-to-tip graph of sampling date of House Finch *Mycoplasma* strains versus divergence from the closest sequence in the putative source population TK_2001. A simple regression gives an estimated substitution rate of 1.45×10^{-5}, consistent with estimates from BEAST. See Text S2 and Text S7 for further information.
doi:10.1371/journal.pgen.1002511.g004

Degredation and apparent functional loss of CRISPR loci in House Finch MG

In some bacterial systems, CRISPRs have a well-recognized function in bacterial immunity and defense against phage, although they may possess additional functions, such as gene regulation [41–44]. We extensively catalogued CRISPR repeats in the House Finch and ancestral poultry strains (Figure 5, Text S8, Table S12). In so doing we observed drastic changes in the CRISPR system between House Finch and poultry strains (Figure 5) [45–48]. The House Finch MG strains from 1994–96 contain up to 50 unique spacers, none of which is shared with the four divergent poultry genomes, which each contained a unique set of 36 to 147 spacer regions consistent with a high rate of turnover for a population actively acquiring new spacer sequences. We found that less than 1% of the 302 unique spacer sequences had similarity to any sequences in the House Finch MG genomes and that none of the remaining spacers had any similarity to sequences in Genbank, indicating an external source for these sequences (Text S8). Surprisingly, no novel spacer elements are present in any of the House Finch MG samples or TK_2001, indicating that the CRISPR array ceased recruiting additional spacers around the time of host switch into the House Finch. In fact, over the 13-year period of the epizootic, the number of unique spacers present in the CRISPR array of the samples decreased to 28 (Figure 5). Further evidence for degradation of the CRISPR locus following the host switch is the complete loss of the four CRISPR-associated (i.e. "CAS") genes in all of the 2007 isolates, a loss that likely renders the CRISPR system in House Finch MG non-functional [45].

Discussion

Rapid substitution rate

We conducted whole-genome sequencing on a unique 13-year serial sample of *Mycoplasma* strains circulating in wild House Finches to characterize genomic changes accompanying a host shift from poultry in the mid-1990s as well as to obtain a very high substitution rate for this avian pathogen. Previous estimates using serial samples and/or the known timing of events presumably tied to the divergence of bacterial strains have generally found much lower rates. An estimate of 2.0×10^{-6} was obtained for *Staphyloccous aureus* [12], 1.1×10^{-7} for *Buchnera* [19], 7.42×10^{-7} in *Yersinia pestis* and 1.4×10^{-6} in *Heliobacter pylori* [14]. Disentangling the effects of recombination and point substitution can be challenging and some previously published substitution rates are likely to be upper bounds rather than point estimates [12]. Our estimate appears to be among the highest reported for a bacterium, and is consistent with other reports of exceptionally high substitution rates in mycoplasmas [37].

Estimates of substitution rates can be influenced by the interval over which sequences are sampled, with estimates taken from short time intervals often exceeding those taken on biogeographic or geological time scales [49]. However the small number of SNPs that we detected segregating in House Finch MG populations suggest negligible effects of multiple hits on our estimate, and our use of a coalescent model suggests that effects of ancestral polymorphism on substitution rate estimates should be adequately accounted for [32,50]. Additionally, our estimates of substitution rate were robust to many potential complicating factors, including

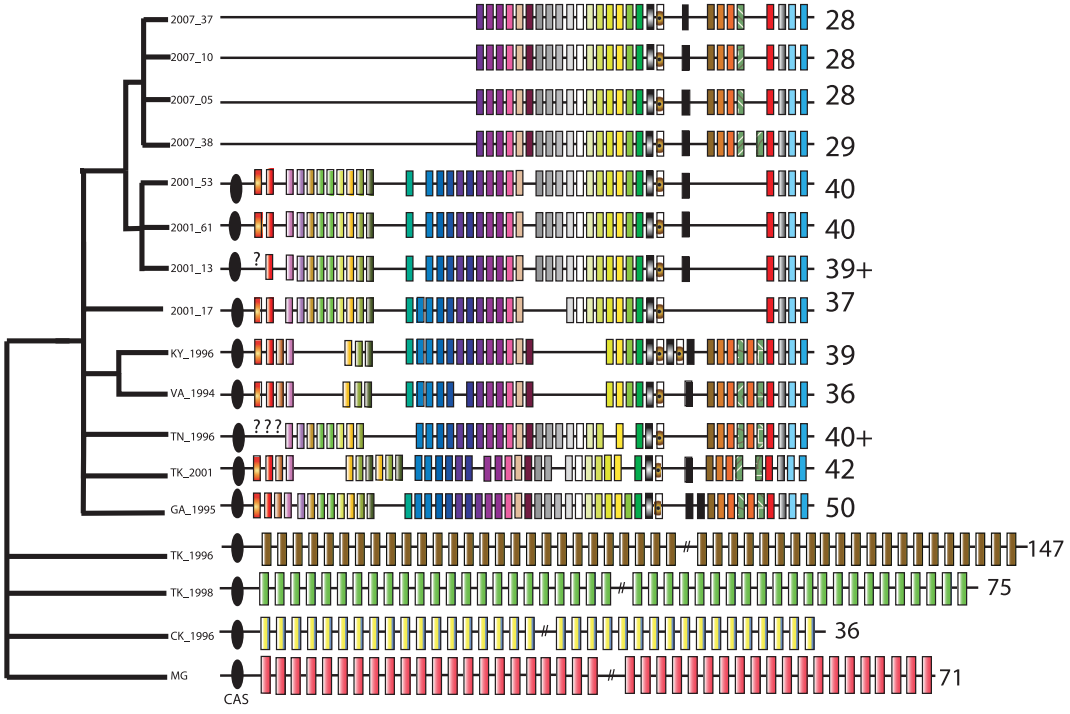


Figure 5. Evolution of the CRISPR locus in *Mycoplasma gallisepticum* isolates collected from House Finches, chickens, and turkeys. Numbers by each strain indicate the number of repeats in each CRISPR array. The ancestral 71-repeat CRISPR array of the chicken MG strain is shown in simplified form at bottom. Diagnostic CRISPR repeats for House Finch MG isolates are indicated in repeat-specific patterns. The black ovals signify the cluster of four CRISP-associated (CAS) genes, which are deleted in the 2007 strains. The tree at left is broadly consistent with the tree based on SNPs (Figure 3) but emphasizes strain clusters indicated by rare genomic changes and CRISPR deletions; it was constructed as described in Text S3.
doi:10.1371/journal.pgen.1002511.g005

SNP calling protocol and whether poultry or House Finches were used as the host for sampled sequences. Given the history and genetic isolation of the House Finch MG strains, the influence of recombination or lateral gene transfer on our estimate of substitution rate is likely also minimized (Text S7).

Rapid evolution and degradation of CRISPRs

The CRISPR dynamics we observed in House Finch MG differ from that seen in other pathogen and bacterial populations. A recent study of *Y.* pestis CRISPR arrays from 131 strains [51] indicated a slower pace of CRISPR evolution than observed in MG and pattern of evolution in which acquisition of novel sequences does not play a prominent role. This study found that in *Y. pestis* the first part of the CRISPR arrays were conserved and that over 76% of all spacer sequences derived from within the *Y. pestis* genome. Similarly, a recent study of *E. coli* and *Salmonella* genomes found that strains within 0.02% divergence typically have identical CRISPR loci [52] and that spacer sequences were often matched to elements of the *E. coli* genome. Additionally, some spacer sequences were shared between strains within a species exhibiting over 1% sequence divergence. These observations and an estimated substitution rate on the order of 10^{-10} per site per year suggested that *E. coli* strains that had diverged for 1,000 years sometimes shared identical CRISPR loci, suggesting patterns of evolution different from that expected for a rapidly changing adaptive immune system primed to combat phages, a conclusion that was supported by later work [53].

By contrast to the pattern seen in these γ-proteobacteria, none of the House Finch MG strains in this study have the same CRISPR locus despite differing at only 0.01–0.02% of sites and likely having last shared a common ancestor less than 20 years ago. Our serial sampling suggests that the loss of spacer sequences and the CRISPR system itself can take place on very short time scales in *Mycoplasma*. Unlike the patterns seen in *E. coli*, *Y. pestis*, and *Salmonella*, the poultry MG strains in our study did not share any spacer sequences, even though they differed by ~1%. These strains had very large CRISPR arrays and 99% of all spacer sequences did not match any known sequence in their genome or in the databases. Therefore the MG CRISPR loci studied here differ from the those observed in some γ-proteobacteria, a group for which CRISPR dynamics can appear functionally unrelated to ecology or immunity [53–55].

Instead, our finding of rapid evolution and degradation of the CRISPR loci more closely resembles patterns found in other bacterial groups, particularly those in which CRISPR is involved in phage defense [56]. CRISPRs are found in only 40% of sequenced bacteria investigated thus far, and often have major roles in bacterial immunity in several lineages investigated in detail [45]. We were surprised to find a gradual degradation and ultimate apparent functional loss of the CRISPR system in House Finch MG after the host switch and a shift in CRISPR dynamics appears to be a major correlate of host switch in this system. One possible explanation for this pattern is that MG experienced release from its ancestral phage parasite community (or other mobile genetic elements such as plasmids) following introduction into the House Finch. Loss of traits upon removal of the agent of selection is a common evolutionary response, as are population expansions of animals and plants when introduced into novel habitats unaccompanied by their parasites [57].

Despite the large amount of ecological research focusing on this host-pathogen system [3–7], at present nothing is known about phages that infect MG or their role in its evolutionary dynamics. Therefore the hypothesis of parasite release as a driver of CRISPR loss is purely speculative. We know of no phage known to infect the Pneumoniae phylogenetic group of mycoplasmas and the few phages known to infect *Mycoplasma* have proven difficult to characterize [58]. We might expect *Mycoplasma* bacteriophages to be host-specific given that they seem to be unusual in their ability to bind to a bacterium with no cell wall and a diverse assortment of surface proteins [58]. However, we are not aware of even basic data on the degree to which *Mycoplasma* might be susceptible to the many bacteriophages that they presumably encounter in their environment. Although phage represent one possible source for these novel ~30 bp sequences, another possible explanation for the source of the spacer sequences is that they derive from plasmids. Although unprecedented (we know of no examples of a naturally occurring plasmid in the Pneumoniae mycoplasmas), such a scenario could raise the possibility of easier genetic manipulations in MG where development of such tools has been challenging [59]. Of the many other possibilities that could explain the observed degradation of the CRISPR loci, we can at least rule out self-interference as an explanation in derived MG strains, given that there is only a single CRISPR cluster in House Finch MG [54]. Measurement of costs, possible advantages and consequences of CRISPR loss, as well as functional and evolutionary assays and surveys of phage diversity will help determine if the rapid and deadly spread of *Mycoplasma* following their expansion into the House Finch was facilitated by a lack of phage predation, a short-term advantage of CRISPR degradation or some other, possibly neutral, mechanism. Although our sequence data is suggestive, explicit functional studies will also be required to demonstrate CRISPR functionality or lack thereof in poultry and House Finch MG and its role, if any, in phage defense.

Pseudogenization and possible mutator strains

Genome evolution of MG during its host-switch from poultry to House Finches adds to a growing list of host-switches that are successful in the complete absence of novel genes [21,60,61] and bacterial lineages exhibiting high rates of point substitution [14]. *Mycoplasmas* are some of the fastest evolving organisms on earth [62] having lost many of the repair mechanisms present in other bacteria [38] and this high mutation rate could help introduce deleterious mutations and contribute to the substantial level of pseudogenization that was observed in this study. The high basal substitution rate in MG may well be elevated yet further by *UvrB* mutation that we detected, a mutation that could have consequences for the long term genomic integrity of this MG lineage, particularly if it remains genetically distinct from and unable to exchange genes with the poultry MG lineages with a functional *UvrB*. Alternatively, given the short (3 amino acid) truncation of this gene in the House Finch strains, another explanation for the greatly increased number of doublet mutations in the lineage carrying the *UvrB* truncation is that selection has not had enough time to remove them as it has for poultry strains without this mutation. Although mutator strains are known to have a selective advantage in rapidly evolving laboratory and natural populations [39,63], additional functional and experimental work will be required to determine the selective and functional effect of the mutation we have detected in *UvrB*, and over what time scales such selective effects might persist. For this and other endeavors, serial sampling of additional bacterial populations in nature will further clarify the rate at which genomes are remolded during host switches in the wild.

Materials and Methods

Sampling of House Finch and poultry MG strain diversity

DNA sequence data for 4 gene fragments collected from 74 strains in Ferguson et. al. [28], was combined with data from 8

strains newly sequenced in this study to yield a Large Sample Multiple Sequence Alignment (LS-MSA) 1,363 bp in length (Figure S2). We estimated nucleotide diversity and the standard deviation of this estimate within and among subgroups of these sequences using DNAsp version 4.10.9 [64] (Table S5). In estimating diversity of MG strains sampled from chickens and turkeys, we restricted analysis to those strains sampled during 1994–1996 for comparison with our earliest House Finch strains sampled in a similar time interval.

Strain selection and genome sequencing

Twelve strains of MG isolated from House Finches in the Southeastern US were sequenced with the Roche 454 Gene Sequencer. The average coverage level was 9.4X (Table S1). Additionally, four MG strains isolated from poultry hosts and selected based on their positions in the multistrain phylogenetic tree were sequenced with the Illumina sequencing platform to an average coverage of ~410 X (Tables S2, S3, S4, Text S1, Figure S2).

Inference of substitutions rates, times to common ancestry, and population dynamics

Using a coalescent model and a Bayesian framework as implement in BEAST v1.52 [32] we estimated the mutation rate and times to common ancestry from a 13-taxon alignment composed of the reference MG genome and all of the House Finch MG strains whose genomes were sequenced in this study (Text S2). We also ensured that the conclusions from this inference were not sensitive to the SNP calling procedures or the choice of substitution models (Text S2, S7, Figure S6). In order to compare the mutation rate between the poultry and House Finch MG populations, these quantities were similarly estimated from the 82 taxon LS-MSA after removing nine laboratory strains from the alignment that likely experienced different population dynamics than the wild strains and had unknown sampling dates. A Poisson regression model was also used to estimate substitution rates by counting mutations along a single lineage assumed to span the dates of sampling for each strain (Text S7).

Transposon movements, recombination, and lateral gene flow

We catalogued IS elements using BLAST and the ISFinder database [65, Text S4]. We tested for evidence of genetic recombination between MG strains using the genome sequences from our 4 poultry and 2 House Finch strains using the pairwise homoplasy index test [34] as implement in splitstree4 [66], and the homoplasy test by Maynard-Smith and Smith [35]. Further evidence for the presence of recombination and the number of nonrecombining blocks was provided by other methods (Text S6, Figures S3, S4, S5).

Supporting Information

Figure S1 To understand the broad phylogenetic diversity of House Finch and poultry MG strains, guide our choice of poultry strains for genomic sequencing and compare mutation rates in the HF and poultry MG population, we used DNA sequence data from Ferguson et al. [28] to generate a multisequence alignment for 82 MG strains collected from four host species (Turkey, Chicken, House Finch and Gold Finch). This data, henceforth the Large Sample Multiple Sequence Alignment, LS-MSA) was composed of four gene fragments (from pvpA, mgc2, gapA and an unnamed surface lipoprotein) that when concatenated yielded approximately 1.9 kb of sequence data per strain (with the exact length of each strain varying due to small indels). We added to this dataset sequences for 8 of the 12 House Finch MG strains sequenced in this study that had complete coverage for these gene fragments. The four strains from this study not incorporated into the dataset (TN_1996, GA_1995, AL_2001_53 and AL_2007_05) were excluded because there was not enough sequencing data to accurately assemble the relevant fragments. We also excluded 3 strains from the original work[28] where we could not identify the host-animal species, leaving 82 strains in the final multiple sequence alignment. In this alignment, all the House Finch haplotypes were identical, except for the 2007 strains that differed from the others at two adjacent nucleotide positions. Certain sections of the gene fragments in the LS-MSA were polymorphic due to insertions/deletions of tandem repeats, and because there is no clear criteria by which to assign the locations of these repeats in an alignment for phylogenetic purposes, for analysis purposes we reduced the ~1.9kb of sequence down to1,36 bp that could be confidently aligned. The tree shown is a phylogeny of 82 avian MG strains inferred from four concatenated gene-segments, totaling 1,363 bp, using Neighbor-joining in PHYLIP. Due to recombination in *Mycoplasma gallisepticum,* this single tree may not be completely representative of the organismal history of the strains from which the gene segments were sampled. However, the pattern showing poultry hosts interspersed amongst the leaves of the tree and high diversity within the MG population is also present in neighbor-joining trees separately inferred for each individual gene fragment, consistent with frequent host-shifts by MG. Strain K4366GF97_10 is from an American Goldfinch (*Carduelis tristis*), also a songbird and the chicken reference strain used to obtain the reference genome is R63_44.
(EPS)

Figure S2 Cross Validation of the 454 Sequencing Data with the Illumina Sequencing Data. Our dataset provides an opportunity to validate the SNP calls made with our 4X-19X coverage 454 data for the House Finch MG isolates by using the SNP calls made with the 294X coverage Illumina data that was generated for TK_2001. TK_2001 and the House Finch MG isolates (particularly the pre-2001 isolates) are nearly genetically identical, and SNPs for both strains were called relative to the much more distantly related strain that was used to generate the reference genome. As outlined with the unrooted tree shown in this figure. This means that most of the SNPs called for each of the House Finch isolates should also be called for the TK_2001 strain, with any unmatched SNPs likely due to either genetic divergence between the two strains or SNP calling errors. The results of this comparison are shown in Table S4. For our most stringent threshold, of the up to 6,461 SNPs that were called in our pre-2001 House Finch isolates, 99.7% of the SNPs called with the 454 data were also called with the Illumina data. This bounds the false positive rate for SNP calls in the 454 stringent data at 0.3%. However, we believe that this unmatched 0.3% is due to true genetic divergence between the strains and not sequencing errors, as these SNPs are very well supported. For example, all 21 SNPs in VA_1994 that did not match TK_2001 were supported by at least 9 reads that contained the variant, and often many more. Table S4 documents the robustness of our population genetic estimates on variations in SNP calling protocol, leading only to minor variations (~1%) in the false positive rate for our SNP datasets. This shows that almost all of the uncertainty in estimating the mutation rate from these genomes is due to the inherent sampling variability that naturally results from the stochastic process that generated them and is not due to any variability that comes from calling SNPs in these genomes. Additionally the ratio of

polymorphic to conserved sites is equivalent across all three datasets.
(EPS)

Figure S3 Illustration of the recursive method used to assign segments of the genome to phylogenetically concordant blocks. At the initialization of the algorithm the phylogenetically informative SNPs in the genome (x's in the diagram) are used to determine continuous segments that are in agreement with all possible trees. Sections of a genome in agreement with a particular tree are shown as solid colored lines over that genome segment. Note that any one SNP can be in agreement with multiple trees. If only one of two adjacent SNPs are in agreement with a tree, then half of the distance between the two SNPs is assigned to the concordant segment.
(EPS)

Figure S4 Distribution of the number of phylogenetically concordant segments in the genome and in a dataset obtained by a single random permutation of the SNPs. Block sizes are in bp.
(EPS)

Figure S5 Distribution of the size of phylogenetically concordant segments in the genome and in a dataset obtained by repeatedly creating permutations of the SNPs.
(EPS)

Figure S6 95% HPD intervals of the rate estimated in BEAST using our actual dataset, as well as 20 permutations of the data where the dates on the tips are randomly reassigned. The interval for the true dataset is shown in red, and the randomized datasets are shown in blue.
(EPS)

Table S1 Characteristics of MG isolates used in this study.
(PDF)

Table S2 SNP counts in the alignments.
(PDF)

Table S3 SNPs validated by PCR amplification and Sanger sequencing.
(PDF)

Table S4 Cross validation of 454 and Illumina data.
(PDF)

Table S5 Estimates of genetic diversity based on the LS-MSA.
(PDF)

Table S6 Patterns of synonymous and nonsynonymous substitutions.
(PDF)

Table S7 Regions of the reference genome that had been lost in House Finch MG isolates.
(PDF)

Table S8 Descriptions of six novel insertion sites of IS elements.
(PDF)

Table S9 Comparative evaluation of genes pseudogenized or deleted in the House Finch MG isolates.
(PDF)

Table S10 Mutations in the *UvrB* gene and possible effects.
(PDF)

Table S11 Instances of polymorphic adjacent SNPs among the house finch MG strains.
(PDF)

Table S12 Counts of unique and total (due to duplication) CRISPR spacers from each strain.
(PDF)

Text S1 Sequencing, alignment, and SNP calls.
(PDF)

Text S2 Inference of mutation rate, recombination, times to common ancestry, and population dynamics.
(PDF)

Text S3 Evaluating the effect of frameshift and nonsense mutations.
(PDF)

Text S4 Transposon (IS) Movements.
(PDF)

Text S5 Searching for Novel Genes in the House Finch MG isolates.
(PDF)

Text S6 Detecting recombination.
(PDF)

Text S7 Effect of recombination on the estimated substitution rate and demonstration of true temporal signal.
(PDF)

Text S8 CRISPR Analysis
(PDF)

Acknowledgments

We thank J. Jones at the University of South Carolina Environmental Genomics Core Faciltiy for conducting the 454 sequencing; the University of Utah Huntsman Cancer Institute for Illumina sequencing; J. Banfield, C. Dale, D. Jeruzalmi, W. Smith, and members of the Marx lab for helpful dicussion; and R. Kofler for sharing programming code libraries. T. Modak and I. Soltero conducted manual SNP verification. S. Bensch, C. Anderson, and three anonymous reviewers provided helpful comments on the manuscript. Sequence data and BEAST xml files for this project were deposited in the Dryad Repository: doi:10.5061/dryad.353tj334.

Author Contributions

Conceived and designed the experiments: SVE GEH NFD. Performed the experiments: NFD CB SB. Analyzed the data: NFD SVE PT AR. Contributed reagents/materials/analysis tools: NFD CJM NF-N. Wrote the paper: NFD SVE GEH AR CJM.

References

1. Fischer J, Stallknecht D, Luttrell P, Dhondt A, Converse K (1997) Mycoplasmal conjunctivitis in wild songbirds: the spread of a new contagious disease in a mobile host population. Emerg Infect Diseases 3: 69.
2. Nolan P, Hill G, Stoehr A (1998) Sex, size, and plumage redness predict house finch survival in an epidemic. Proceedings of the Royal Society B-Biological Sciences 265: 961.
3. Dhondt AA, Dhondt KV, Hawley DM, Jennelle CS (2007) Experimental evidence for transmission of *Mycoplasma gallisepticum* in house finches by fomites. Avian Pathol 36: 205–208.
4. Dhondt AA, Tessglia DL, Slothower RL (1998) Epidemic mycoplasmal conjunctivitis in House Finches from eastern North America. J Wildlife Dis 34: 265–280.
5. Faustino C, Jennelle C, Connolly V, Davis A, Swarthout E, et al. (2004) *Mycoplasma gallisepticum* infection dynamics in a house finch population: seasonal variation in survival, encounter and transmission rate. Ecology 73: 651–669.
6. Hochachka WM, Dhondt AA (2000) Density-dependent decline of host abundance resulting from a new infectious disease. Proc Natl Acad Sci (USA) 97: 5303–5306.

7. Luttrell M, Fischer J, Stallknecht D, Kleven S (1996) Field investigation of *Mycoplasma gallisepticum* infections in house finches (*Carpodacus mexicanus*) from Maryland and Georgia. Avian Dis. pp 335–341.
8. Rambaut A, Pybus O, Nelson M, Viboud C, Taubenberger J, et al. (2008) The genomic and epidemiological dynamics of human influenza A virus. Nature 453: 615–619.
9. Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA (2007) A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc Natl Acad Sci (USA) 104: 7993–7998.
10. Barrick J, Yu D, Yoon S, Jeong H, Oh T, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243–1247.
11. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469–474.
12. Nubel U, Dordel J, Kurt K, Strommenger B, Westh H, et al. (2010) A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant *Staphylococcus aureus*. PLoS Path 6: e1000855. doi:10.1371/journal.ppat.1000855.
13. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, et al. (2011) Rapid Pneumococcal Evolution in Response to Clinical Interventions. Science 331: 430–434.
14. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, et al. (2010) Microevolution of Helicobacter pylori during Prolonged Infection of Single Hosts and within Families. PLoS Genet 6: e1001036. doi:10.1371/journal.pgen.1001036.
15. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, et al. (2010) Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc Natl Acad Sci (USA) 107: 7527–7532.
16. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet 40: 987–993.
17. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, et al. (2006) Evolutionary history of Salmonella Typhi. Science 314: 1301–1304.
18. Morelli G, Song YJ, Mazzoni CJ, Eppinger M, Roumagnac P, et al. (2010) Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42: 1140–1143.
19. Moran N, McLaughlin H, Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323: 379.
20. Ochman H, Elwyn S, Moran N (1999) Calibrating bacterial evolution. Proc Natl Acad Sci (USA) 96: 12638–12643.
21. Parkhill J, Sebaihia M, Preston A, Murphy L, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35: 32–40.
22. Eppinger M, Baar C, Linz B, Raddatz G, Lanz C, et al. (2006) Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. PLoS Genet 2: e120. doi:10.1371/journal.pgen.0020120.
23. Papazisi L, Gorton TS, Kutish G, Markham PF, Browning GF, et al. (2003) The complete genome sequence of the avian pathogen Mycoplasma gallisepticum strain Rlow. Microbiol 149: 2307–2316.
24. Farmer KL, Hill GE, Roberts SR (2002) Susceptibility of a naive population of house finches to Mycoplasma gallisepticum. J Wildlife Dis 38: 282–286.
25. Nolan PM, Roberts SR, Hill GE (2004) Effects of Mycoplasma gallisepticum on reproductive success in house finches. Avian Dis 48: 879–885.
26. Wang Z, Baker AJ, Hill GE, Edwards SV (2003) Reconciling actual and inferred population histories in the house finch (Carpodacus mexicanus) by AFLP analysis. Evolution 57: 2852–2864.
27. Wang Z, Farmer K, Hill GE, Edwards SV (2006) A cDNA macroarray approach to parasite-induced gene expression changes in a songbird host: genetic response of house finches to experimental infection by Mycoplasma gallisepticum. Mol Ecol 15: 1263–1273.
28. Ferguson N, Hepp D, Sun S, Ikuta N, Levisohn S, et al. (2005) Use of molecular diversity of *Mycoplasma gallisepticum* by gene-targeted sequencing (GTS) and random amplified polymorphic DNA (RAPD) analysis for epidemiological studies. Microbiol 151: 1883–1893.
29. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, et al. (2010) Multi-Platform Next-Generation Sequencing of the Domestic Turkey (Meleagris gallopavo): Genome Assembly and Analysis. PLoS Biol 8: e1000475. doi:10.1371/journal.pbio.1000475.
30. Dimcheff DE, Drovetski SV, Mindell DP (2002) Phylogeny of Tetraoninae and other galliform birds using mitochondrial 12S and ND2 genes. Mol Phyl Evol 24: 203–215.
31. Barker FK, Cibois A, Schikler P, Feinstein J, Cracraft J (2004) Phylogeny and diversification of the largest avian radiation. Proc Natl Acad Sci (USA) 101: 11040–11045.
32. Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
33. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution 22: 1185–1192.
34. Bruen T, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665.
35. Maynard Smith J, Smith N (1998) Detecting recombination from gene trees. Mol Biol Evol 15: 590.
36. Posada D, Crandall K (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America 98: 13757.
37. Woese C, Stackebrandt E, Ludwig W (1985) What are mycoplasmas: the relationship of tempo and mode in bacterial evolution. J Mol Evol 21: 305–316.
38. Carvalho F, Fonseca M, Batistuzzo De Medeiros S, Scortecci K, Blaha C, et al. (2005) DNA repair in reduced genome: the mycoplasma model. Gene 360: 111–119.
39. Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of E. coli. Nature (London) 387: 703–705.
40. Moolenaar G, Franken K, Dijkstra D, Thomas-Oates J, Visse R, et al. (1995) The C-terminal region of the UvrB protein of Escherichia coli contains an important determinant for UvrC binding to the preincision complex but not the catalytic site for 3-incision. Journal of Biological Chemistry 270: 30508.
41. Levin BR (2010) Nasty Viruses, Costly Plasmids, Population Dynamics, and the Conditions for Establishing and Maintaining CRISPR-Mediated Adaptive Immunity in Bacteria. PLoS Genet 6: e1001171. doi:10.1371/journal.pgen.1001171.
42. Nozawa T, Furukawa N, Aikawa C, Watanabe T, Haobam B, et al. (2011) CRISPR Inhibition of Prophage Acquisition in Streptococcus pyogenes. PLoS ONE 6: e19543. doi:10.1371/journal.pone.0019543.
43. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6: 181–186.
44. Vale PF, Little TJ (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. Proceedings of the Royal Society B-Biological Sciences 277: 2097–2103.
45. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Genetics 6: 181–186.
46. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709–1712.
47. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, et al. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190: 1390–1400.
48. Tyson G, Banfield J (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Envir Microbiol 10: 200–207.
49. Ho SY, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22: 1561–1568.
50. Emerson BC (2007) Alarm bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. Syst Biol 56: 337–345.
51. Cui Y, Li Y, Gorgé O, Platonov ME, Yan Y, et al. (2008) Insight into microevolution of Yersinia pestis by clustered regularly interspaced short palindromic repeats. PLoS ONE 3: e2652. doi:10.1371/journal.pone.0002652.
52. Touchon M, Rocha EPC (2010) The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. PLoS ONE 5: e11126. doi:10.1371/journal.pone.0008694.
53. Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, et al. (2011) CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. J Bacteriol 193: 2460–2467.
54. Diez-Villasenor C, Almendros C, Garcia-Martinez J, Mojica FJ (2010) Diversity of CRISPR loci in Escherichia coli. Microbiol 156: 1351–1361.
55. Cady KC, White AS, Hammond JH, Abendroth MD, Karthikeyan RS, et al. (2011) Prevalence, conservation and functional analysis of Yersinia and Escherichia CRISPR regions in clinical Pseudomonas aeruginosa isolates. Microbiol 157: 430–437.
56. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327: 167.
57. Torchin M, Lafferty K, Dobson A, McKenzie V, Kuris A (2003) Introduced species and their missing parasites. Nature 421: 628–630.
58. Waldor MK (2005) Phages: their role in bacterial pathogenesis and biotechnology; Waldor MK, Friedman DI, Adhya SL, editors. Washington, ed. D.C.: American Society of Microbiology Press.
59. Lee SW, Browning G, Markham P (2008) Development of a replicable oriC plasmid for Mycoplasma gallisepticum and Mycoplasma imitans, and gene disruption through homologous recombination in M. gallisepticum. Microbiol 154: 2571.
60. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.
61. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MTG, et al. (2001) Genome sequence of Yersinia pestis, the causative agent of plague. Nature 413: 523–527.
62. Ciccarelli F, Doerks T, Von Mering C, Creevey C, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283.
63. Hoboth C, Hoffmann R, Eichner A, Henke C, Schmoldt S, et al. (2009) Dynamics of adaptive microevolution of hypermutable Pseudomonas aeruginosa during chronic pulmonary infection in patients with cystic fibrosis. J Infect Disease 200: 118.

64. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496–2497.
65. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucl Acids Res 34: D32–D36.
66. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23: 254–267.
67. Hillier LDW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nature Methods 5: 183–188.
68. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Research 18: 763.
69. Martin D (2009) Recombination detection and analysis using RDP3. Methods Mol Biol 537: 185–205.
70. Jolley K, Feil E, Chan MS, Maiden MCJ (2001) Sequence type analysis and recombinational tests (START). Bioinformatics 17: 1230.
71. Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML.
72. Duffy S, Holmes EC (2009) Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. Journal of General Virology 90: 1539.
73. Ley D, Berkhoff J, Levisohn S (1997) Molecular epidemiologic investigations of Mycoplasma gallisepticum conjunctivitis in songbirds by random amplified polymorphic DNA analyses. Emerging Infectious Diseases 3: n3.
74. Ley D, Berkhoff J, McLaren J (1996) *Mycoplasma gallisepticum* isolated from house finches (*Carpodacus mexicanus*) with conjunctivitis. Avian Dis. pp 480–483.
75. Tully JG, Razin S (1983) Diagnostic mycoplasmology. New York: Academic Press. xxiii, 440 p. p.
76. Farmer K, Hill G, Roberts S (2005) Susceptibility of wild songbirds to the house finch strain of Mycoplasma gallisepticum. J Wildlife Dis 41: 317.
77. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115. doi:10.1371/journal.pgen.1001115.
78. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24: 1586–1591.
79. Barre A, de Daruvar A, Blanchard A (2004) MolliGen, a database dedicated to the comparative genomics of Mollicutes. Nucleic Acids Research 32: D307–310.
80. Moolenaar G, Franken K, van de Putte P, Goosen N (1997) Function of the homologous regions of the Escherichia coli DNA excision repair proteins UvrB and UvrC in stabilization of the UvrBC–DNA complex and in 3 -incision. Mutation Research-DNA Repair 385: 195–203.

Text S1: Sequencing, Alignment and SNP Calls

We used a mixed 454 and Illumina sequencing strategy. Genomic DNA from each House Finch strain was first nebulized and blunt-end ligated with Roche adaptors. Emulsion PCR was performed using the bulk library and these products were pyrosequenced on Roche 454 Gene Sequencer using FLX chemistry, using physical separation of isolates. Genomic DNA from the four newly sequenced poulty strains (Table S1) was sequenced by the Illumina method at the University of Utah Huntsman Cancer Institute.

Sequence data generated from the Illumina sequencing platform was used to make a multiple sequence alignment according to the following protocol. Raw sequences were first trimmed to retain only high quality sequence data. Trimmed reads over 25 bp in length were then aligned to the reference MG genome (AE015450.2), ignoring any ambiguously mapped reads using the CLC Genomics Workbench version 3.7.1 . Finally, bases were called from the consensus sequence in this alignment for each unmasked (see below) regions of the genome. Any basepair in these regions that differed from the consensus region was only called as a SNP if there were at least 4 reads at the position, if the base passed NQS (30/25) standards, and if at least 95% of all reads aligning at that position had the SNP base. To avoid errors due to runs of single bases, we also required that there be no more than 2 gaps or mismatches within an 11 bp window around the putative SNP in any read that was counted towards calling the SNP. This final criterion is effective at removing erroneous SNPs that are simply artifacts of poor sequence alignment. However, due to the high number of SNPs present in all four of our sequenced strains (10,007-10,729 differences in an alignment of ~756 kb), we would expect this to also exclude some legitimate SNPs that fell in windows with 2 or more neighboring SNPs simply by chance. For this reason, we generated a second dataset that varied this criterion by allowing up to 4 gaps or mismatches with in an 11 bp window around any position that differed from the reference genome. This dataset contained only 1% more SNPs than the original dataset, and we found that the conclusions in this paper were qualitatively unaffected by using this alternate dataset.

The sequencing data from the House Finch MG isolates generated on the Roche 454 platform data was aligned to the reference *Mycoplasma gallisepticum* genome [1] using the Mosaik aligner [2]. From this alignment a basepair for a strain was reported for each position in the reference genome if the following conditions were met. The majority base at that position had to have a center QC score of 30 or higher, and the 5 flanking bases on either side of it to each have a score of 25 or higher. We also required at least two reads before a base was called. A whole genome alignment was then generated by aligning the basepairs present from each sample at the same position in the reference genome. This whole genome alignment was then divided into sections and the alignments were manually curated by the authors. If during this process a region of the genome was found to clearly violate the Markov models we assumed for nucleotide evolution (e.g. a transposon insertion) or if the aligner was grossly in error (e.g. a section with a deletion or a slightly varying simple sequence repeat that confused the aligner) we either manually edited the relevant section or if it was not obvious what the correct alignment should be we excluded it from later analysis. We additionally validated our alignments by ensuring that the base we reported at every position in our alignments matched the base independently reported by a separate alignment algorithm. The other aligner we used was the run454Mapper program (part of the Genome Sequence FLX Data Analysis Software package available from Roche).

For both the Illumina and 454 sequencing data, we aggressively excluded repetitive segments of the genome that we believed to be inappropriate for SNP calls. In particular, the *Mycoplasma gallisepticum* genome contains a number of proteins that have a high degree of similarity with other proteins, such as the VlhA family of lipoproteins that constitute 10.4% of the reference genome, and also the Apr-E like proteins, transposases, CRISPRs, etc. This repetitive DNA is likely to undergo recombination and in some cases it is not possible to correctly align or assemble. To avoid artifacts introduced by these regions, we excluded any region of the genome over 100 bp in size that had over 85% similarity to another location in the reference genome as determined by megablast . In total, we excluded 228,875 bases (~23% of the reference, henceforth the masked segments) from our multi-isolate alignment and from our SNP calling protocol.

Alternate SNP Calling Protocols

To check the sensitivity of our results to our SNP calling method for the House Finch MG isolate data, we generated three alternate genomic alignments using different protocols and quality threshold levels. These datasets were all generated for only the unmasked portions of our genome, and are described below. All analyzes described in this paper were also performed with these alternate datasets and equivalent results were obtained.

a) Stringent Threshold - Broad Institute 454Swap SNP Calling Software

This dataset calls SNPs using software developed at the Broad Institute for 454 data [3]. The pipeline that uses this SNP calling software is completely independent of the other base calling methods we used. The quality threshold requirements for this program include:

a) A basepair needs to have reads that align to it coming from both the right and the left side.
b) A basepair must be represented by at least two reads that pass NQS thresholds.
c) No more than 33% of the reads at a position can disagree with the consensus base.

b) Moderately Stringent Threshold – Our working data set

This is our working dataset and was created as described in the preceding section. Using the protocol in which 2 gaps or mismatches within an 11 bp window were removed produced an alignment that was 756,552 bp in length. In addition, we also produced an alignment allowing up to 4 gaps or mismatches within an 11 bp window (with the level of diversity present in our poultry sample this many mutations could be expected to occasionally occur and so this may not always indicate sequencing errors). This alternate alignment of 756,574 bp contained only 1% more SNPs than the original method and use of this alignment did not affect any of the conclusions in the paper. This protocol also yielded an alignment of 738,209 bp when considering only the House Finch MG strains.

c) Moderate threshold -A variant of the moderately stringent data set

This dataset was produced exactly as above except there was no requirement that the data generated matched the data generated by the Roche 454 aligner.

Further SNP Dataset Validation

In addition to the checks described above, we also performed 76 traditional Sanger sequencing reactions of SNPs called in the House Finch MG dataset (Table S3) and confirmed that the SNPs were called correctly. Additionally, as our dataset is a composite of 454 and very high coverage Illumina sequencing data, we cross-validated our results by comparing these two datasets to each other and found excellent agreement between them (Figure S2, Table S4).

Text S2: Inference of mutation rate, recombination, times to common ancestry and population dynamics

Using BEAST v1.52 [4], estimates of times of common ancestors were obtained for both the 13-taxon alignment of 738 kb containing our 12 House Finch Strains and the reference genome (large alignment) as well as for 73-taxon LS-MSA of 1.3 kb, which included MG sequence data obtained from strains sampled between 1955 and 2000 [5]. To aid in the selection of the inference model and to ensure that the results based on the large alignment were qualitatively insensitive to inference model choice, we compared the estimates of the mutation rate obtained from a variety of different possible analyses. Since a population expansion was observed to occur over the sampling period, in all inference models considered we assumed a changing population size using the exponential skyline model [6], and also always assumed some form of the HKY nucleotide substitution model. Given these model choices, we also tested the effect of four additional choices, or factors, on our inference. One of these factors was the modeling choice for site heterogeneity, which we tried at three levels (HKY, HKY+Γ or HKY+Γ+I). We also varied the data by including and excluding the reference genome because it was sampled at a much earlier time point then the other strains and thus could exert a high amount of leverage on the rate estimate. Another factor was the multiple sequence alignment used, and we tested all three of our SNP calling datasets (Stringent, Moderate and Moderately Stringent). Finally, since the amount of sequencing data present for each of our strains varied, we tested whether the strains with greater coverage were biasing the results by running the analysis while allowing BEAST to average over partially observed sites, or by only analyzing sites with data for all strains.

In total this resulted in 36 ($3 \cdot 2 \cdot 3 \cdot 2 = 36$) different methods to infer the rate of evolution, and inference about the posterior distribution of the rate parameter was obtained for each of these methods from 10,000,000 MCMC samples. From this analysis, 2 of the 36 MCMC runs were unable to converge. These runs were performed with settings that essentially deprived the inference method of enough data to jointly infer the parameters in the model (e.g. the stringent dataset and the requirement that all strains have data present), and as a result the estimates were wildly varying and inaccurate (e.g. Median clock rates of 1.53e307) and the MCMC chains clearly failed to converge. A plot of the rate estimates from the 34 runs that could produce sensible results is shown in Fig. S4.

We concluded from this analysis that the rate estimate was robust to these model choices. However, the results reported in this paper are based on the model we believed to be the best, which included the reference strain to allow inference about its divergence time from the HF ancestor and, used the HKY+I model of substitution (when inferred, the posterior distribution of the Gamma parameter was identical to the prior because the low amount of diversity in the house finch MG meant there were not enough multiple mutations to estimate this parameter), and used our Moderately-Stringent dataset. For this model, we ran 8 additional chains starting from different initial trees and parameter settings, and checked that all converged to the same distribution. The results for this analysis gave an estimated mean clock rate of 1.02e-5 per year (95% HPD 7.95e-6 to 1.23e-5), an estimated date for the MRCA of the HF strains as having lived 19.2 (95% HPD 16.9 to 21.7) years prior to 2007 and estimates the common ancestor of the HF strains and the chicken reference to have occurred 599.2 (95% HPD 477.5 to 737.0) years

prior to 2007. We also used this analysis to estimate a skyline plot for the House Finch MG [6] (Figure 2).

In order to compare our rate estimates with the 73 taxon, 1.3kb alignment, we also estimated these quantities using BEAST, again from 8 different initial values, assuming a model of population change and using the HKY+G+I substitution model. This estimated the mean rate as 3.23e-5 (95% HPD 6.37e-6 to 6.239e-5), and the common ancestor of the HF strains and HF strains to have lived 456.7 (95% HPD 130.8 to 969.4) years prior to 2007. We caution that the estimates of the divergence dates from HF to poultry strains are very coarse and should be interpreted with caution, as the modern poultry industry likely alters the population dynamics of MG transmission in ways that may strongly violate the coalescent model assumed in BEAST.

Text S3: Evaluating the effect of frameshift and nonsense mutations

We looked for frameshift and nonsense mutations, which we refer to collectively as disrupter mutations. To find these mutations, we *de novo* assembled the 454 reads from our House Finch MG samples, and searched for proteins in the assemblies that had such mutations in them. As the 454 *de novo* assembler improves with increasing read coverage, we restricted our analysis to two of our samples with high sequencing coverage, AL_2007_37 and VA_1994. These strains also bookend the time period of this study. Because the TK_2001 strain is so genetically similar to the MG strains in this study isolated from the House Finches, we also searched assemblies generated from its sequencing data, using the CLC genomics workbench v.3.7.1. For all of these strains, we searched for disrupter mutations present in any of the genes annotated in the reference genome, except for genes with strong similarity to other parts of the genome as these genes are most likely to be misidentified or misassembled. We excluded any gene that was annotated as a VlhA gene or a transposon, or that contained a sequence over 100 bp in length that aligned to another area of the genome with over 85% identity as determined by megablast. By this method 105 of 763 genes (13.7%) were excluded.

For each gene of the remaining 658 genes we used the *de novo* reconstructed gene sequences to check for the presence of disruptor mutations. We considered a gene successfully reconstructed if we were able to find a matching segment amongst the assembled contigs that covered the entire gene (as determined by evaluating local alignments determined by Megablast), and that did not differ by more than 200 bp in size. We were able to find matches for all but 48 of the 658 genes (~93% recovery) in our VA_1994 strain, all but 41 in AL_2007_37 (~94%) and all but 20 (97%) in the TK_2001 strain. 17 of the genes were not recovered in VA_1994 and TK_2001 because they had been deleted along the branch leading from the reference MG strain to our isolates, while such deletions caused 29 genes to be unrecoverable in AL_2007_37. The remaining genes were excluded either because they were not completely covered by a single assembled contig, or in one case because an IS element was inserted into it.

To detect pseudogenizing mutations, each of the 617(AL) , 610 (VA) and 622 (TK) successfully reconstructed genes was translated to detect nonsense or frameshift mutations. This identified 85 possible mutations affecting 76 genes in AL_2007_37 and 99 possible mutations affecting 91 genes in VA_1994. For each of these mutations, we then examined the reads supporting them by evaluating the alignment of the reads to the reference genome in the .ace file produced by both the Newbler and Mosaik aligners. We found that many of the indel mutations were near homopolymers where the underlying reads often both supported and contradicted the presence of the relevant indel mutation. We disregarded all such ambiguous cases unless the reads supporting the presence of the indel outnumbered those contradicting it by 10. This criterion excluded 55 mutations in VA_1994 and 41 mutations in AL_2007_37. This left 44 mutations affecting 42 genes in AL_2007_37 and 44 affecting 43 genes in VA_1994. All of these mutations were shared between VA_1994 and AL_2007_37, except for two. One putative nonsense mutation along the branch leading to AL_2007_37 (reference position: 30,546) was found to have occurred in a gene that had already suffered a frameshift in the common ancestor

of VA_1994 and AL_2007_37. A second mutation was present in VA_1994, but because this area of the genome had been deleted in AL_2007_37, it could not be recovered from this sample.

Of the 45 disruptor mutations found, we excluded an additional 18 mutations because the mutation either occurred in a gene that had been annotated as a pseudogene, or because the mutation was actually supposed to be the wild type state of the gene. The later are likely due to sequencing errors or mutations in the reference genome and we determined this to be the case if the effect of the mutation was to merge two pseudogenes back into a functional protein, and if the mutation was present in all of our sequenced poultry strains as well. This left a total of 27 total disruptor mutations which we grouped into the following two categories. All mutations present in the VA_1994 strain were also present in the closely related TK_2001 strain, and some were present in the other poultry strains.

a) Extension Mutations

4 frameshift mutations had the effect of simply extending the length of the protein shown below. These mutations all occurred within the last 1% along the length of the protein, and although these changes do alter the amino acids towards the end of the protein, it is likely that these proteins remain functional.

Genes with extension mutations

Protein ID	Mutation Location	Mutation	Length of extension (aa)	Present in All Strains?
MGA_0809	132,809	A deleted	5	Yes
MGA_0812	135,135	T->A interrupts stop codon	5	Yes
MGA_1153	416,101	Single T deletion in TK_2001, AL and VA have a deletion of 2 "T"'s at this location	2	Only House Finch MG strains and TK_2001
MGA_0232	718,459	A deleted	11	All but TN_1996

b) Pseudogenes Formers

Excluding the extension mutations and mutations that disrupted the reading frame in one gene but merged it with an upstream coding sequence, we observed 23 mutations affecting 17 genes. These were distributed as 10 insertions, 10 deletions and 3 mutations of an amino acid coding codon to a stop codon. The mutations were often clustered in the same gene. There are 4 genes each of which had 2 mutations which would have disrupted the original reading frame, as well as 1 gene with 3 disruptor mutations. The remaining 12 genes were only disrupted by one mutation. The genes affected by these mutations are given in table S9.

Text S4: Transposon (IS) Movements

To identify areas of transposon insertion, and to determine if our isolates contained transposable elements in the same location as the reference genome, we developed a method to identify and annotate transposable elements from the 454 reads. Briefly, the method uses a querying strategy similar to BLAST to search for reads that contain sequences identified with the edges of IS elements. The method then annotates the portion of the read that belongs to the IS element, and maps the remaining portion of the read back to the reference genome in order to identify the location where the IS abuts a portion of the genome. The source code for the method is available from the authors upon request.

The reference genome contains members of 2 groups of IS elements. The first group present is identified by the ISFinder database [16] as belonging to the IS1634 family, and is represented in the reference genome by two complete transposases and one shorter fragment with high similarity to a complete IS element (we refer to such fragments as a scar). The second group includes members of the larger IS256 family, and is represented by ten transposases in the genome (although one of these is broken apart by another copy of an IS which has been inserted into it).

The transposases belonging to the first group (IS1634) have also been found in the genomes of other *Mycoplasma* species including *bovis*, *mycoides*, *hyopneumoniae* and *synoviae* [16]. Although this transposase seems to effectively persist in these other *Mycoplasma* genomes, it appears that no functional copy of this transposase remains in this study's House Finch MG strains. Of the two transposases annotated in the reference genome, only one was functional as the other had a frameshift mutation in it. Based on the Newbler assemblies of our sequence data, this particular transposase is even more degraded amongst the strains we sequenced. The first stop codon now appears only 30 amino acids into the gene in all the strains where we could confidently reconstruct it. The only remaining member of the family present in the reference genome, with the only functional transposase, is gone entirely from the House Finch samples we sequenced. It appears that this remaining functional transposase recombined with one of its scars, leading to a large deletion and the destruction of this last functional copy.

In contrast, the second group of transposases, belonging to the IS256 family, has been active during the divergence from the most recent ancestor of our samples and the reference genome. In the reference genome, this group is represented by 10 transposases and 3 small scars. However, in our samples, only 4 of these 10 IS elements are present. Three IS elements in the reference genome had not been inserted by the time the reference strain and the strains in our samples diverged, and three of the other IS elements were located in a region of the genome that had been deleted in the lineage leading to the common ancestor of all of our samples.

Along the branch leading to the common ancestor of all our samples, this element inserted itself into 6 new locations (Table S8). Each of these insertions shown was present in every one of our HF samples, and no sample had any insertion that was not present in the others. Of the 6 IS element insertions, 4 were in intergenic regions, which given the density of genes in the reference genome is highly unlikely ($p < 0.003$). A likely explanation for this bias is that selection is filtering out insertions that destroy functioning genes.

Text S5: Searching for Novel Genes in the House Finch MG isolates

This study relied on comparing the assembled genomes of the 12 House Finch MG isolates with that of an annotated reference strain. Given the amount of divergence between the reference and our samples, it was important to determine if using this reference genome would prevent us from analyzing additional gene sequences that were not present in the reference genome but that could provide additional information for this study. To investigate the presence of potentially novel genes in our House Finch isolates, we searched the contigs generated via *de novo* assembly for DNA sequences that could not be mapped back to the reference genome. To do this, we megablasted all of the assembled contigs against the reference genome, and examined any section of a contig sequence longer than 100 bp that could not be mapped to the reference genome. To maximize our chances of detecting any novel sequences in the assembled contigs, we examined the contigs generated from our high-coverage VA_1994 and AL_2007_37 strains. We also pooled all of our 2007 samples for *de novo* assembly and investigated the contigs that were generated from this meta-sample.

Few if any novel DNA sequences were found and arguably none were truly unique because they all had strong similarities to members of either the VlhA or AprE-like proteins present in the reference genome. Of the sequences that failed to align with high similarity to the reference genome, several sequence segments ranging in size from 100bp to 2.1 kb could be identified as similar to a VlhA region by BLAST or BLASTX. However, the largest segment of these that aligned with less than 80% similarity to a portion of the reference genome was only 1.6 kb in size, and a translation of this sequence revealed that it contained a Vlh-A type gene. Similarly, there was a ~500 bp segment that could not be mapped to the reference genome, but this segment was flanked by ~3.3 kb of DNA sequence that had between 66-70% similarity with the other AprE-like proteins present in the genome. These results were consistent for all of the assemblies tested. Given the difficulty in reconstructing these repeat-rich loci and their unsuitability for calling SNPs, we did not pursue these segments further.

Text S6: Detecting recombination

Despite the small amount of genetic variation segregating amongst our House Finch *Mycoplasma* samples (only 412 SNPs), it is not possible to build a single phylogenetic tree with no homoplasies from this data. Similarly, although our poultry strains contained many more SNPs between them, one still cannot infer a single phylogenetic tree that has much more support than alternate trees. The reason for this is not that the SNPs provide very little information about phylogenetic relationships, but rather that many SNPs provide information that is in conflict with the information provided by many other SNPs as determined by the four gamete test. This type of behavior is expected if genes are flowing horizontally as well as vertically in a population, and so we formally tested for the presence of recombination in our dataset.

A plethora of tests are available to detect recombination in sequence data (see ref [7-9]). However, because many of these tests examine the same fundamental signal of recombination, such as the physical clustering of phylogenetically concordant SNPs, they commonly yield qualitatively similar results when performed on the same dataset. To detect recombination in our combined House Finch and poultry strain dataset, we used the pairwise homoplasy index test[7] as implement in splitstree4 [10]. Examining the entire data set, this test found a statistically significant signal of recombination ($p < 1e\text{-}9$). This signal comes predominantly from the four newly sequenced poultry strains because there is not enough genetic variation to make the test significant when only the house finch strains are considered. However if we apply to the house finch MG strains the homoplasy test by Maynard-Smith and Smith [11], which is found to perform well in situations of low nucleotide diversity [9], we still obtain a significant signal for recombination. This test differs from the pairwise homoplasy index test, in that rather than looking for spatial clustering of phylogenetically concordant SNPs, it instead asks if the number of homoplasies observed on a tree is particularly large given the number of mutations on the tree and the number of sites that were available to mutate. To implement this test, using the dnapars program in the phylip package we first found a parsimonious tree for the House Finch isolates while including the reference genome as an outgroup, and then counted the number of homoplasies that appear only within the clade of House Finch MG isolates. This identified 13 homoplastic mutations out of a total of 412 variable sites in an alignment of approximately 756.5 kb of DNA. Intuitively, it seems extremely unlikely to see so many homoplastic mutations given the large number of sites that were available to mutate. However, exactly quantifying how unlikely this is complicated because different sites in the genome evolve at different rates, so that homoplasies are much more likely to appear at some sites than others. To get around this issue, the original paper describing the test proposed reducing the total number of sites in the alignment down to a smaller number of effective sites. This paper described a heuristic method to estimate how much one should reduce the alignment size, but this method required a sequence from a distant outgroup that fulfilled a difficult set of assumptions. In practice, since having such an outgroup and trusting that it satisfies the assumptions is rare, many researchers simply take the effective number of sites to be equal to 0.6 multiplied by the total number of sites in the alignment. This 0.6 value was selected as a conservative choice when it was first used in a paper comparing different methods of testing for recombination [9] because it was much less than the inferred values in the original paper (0.73-0.83) and because 0.6 was given as the lower limit for a believable estimate in that same paper during a discussion of different methods to estimate the effective number of sites. Since then, this 0.6 value has been used widely in other papers and is

the default setting implemented in software packages that implement tests for recombination such as START (Sequence Type Analysis and Recombinational Tests, [12]). Suffice is to say, it is clear that there is some ambiguity in how best to determine the effective number of sites that are available to mutate, particularly when there is not complete sequencing for every strain, and as a result the homoplasy test could be considered overly conservative or subjective depending on one's prior beliefs. However, because the observed number of homoplasies in our dataset is so unlikely, we can confirm that it is extremely unlikely even given a wildly conservative set of assumptions. To determine the effective number of sites we used in our test, we first dropped the total number of sites in the alignment from 756,552 bp, down to the total number of sites where all the strains had data present which was only 273, 482 bp. This is obviously an overly conservative reduction as the vast majority of sites in the alignment had data for a majority of strains. Next, instead of applying the standard 0.6 correction to this reduced number, we applied a much more stringent criterion of 0.2, leaving us with 0.2 x 273482 = 54,695 effective sites, or only 7% of the original alignment length. We then estimated the probability of observing 13 or more homoplasies by simulation. Of 1 million simulations, the highest observed number of homoplasies was only 9, and we thus estimated our p-value as $p<1e\text{-}6$. However, the probability of observing so many homoplasies is almost certainly lower than this bound, not only because every assumption we made is expected to increase the p-value, but also because in our dataset two sites needed to convergently mutate not twice, but three times each in order for them to be in agreement with the tree. Since the homoplasy test treats all homoplasy counts as equivalent, even though repeated homoplasies at the same site are particularly unlikely, this again introduces a conservative bias into the test. We also note that the homoplasy test does not consider that a mutation at a site need not produce the exact same basepair each time, as there are three basepairs available to mutate to, which introduces yet one more conservative bias into the test.

Having established that *Mycoplasma gallisepticum* is a bacterium that recombines, we next sought to characterize the nature of recombination in this organism. To bookend a continuum with a dichotomy, recombination between microorganisms can be described as either chunky or smooth. Recombination is chunky when the recombination rate is much lower than the mutation rate, so that the genome is filled with large blocks of easily identifiable DNA that have a shared history that is in strong disagreement with the phylogenetic pattern exhibited by other sections of the genome. In contrast, recombination is smooth when the recombination rate is nearly equal to or greater than the mutation rate, in which case clusters of phylogenetically concordant SNPs tend to be much smaller and correctly delineating a specific section of DNA that has not recombined since the last common ancestor is impossible to do with any reasonable certainty. We therefore looked at the size distribution of phylogenetically concordant chunks to examine the extent to which the statistically significant finding of recombination was due to a few large blocks, or many smaller blocks.

To do this, we systematically determined the size distribution of phylogenetically concordant genomic segments in our sequenced isolates by implementing a recursive method that assigned each possible basepair in the genome to a phylogenetically concordant segment. Our method, illustrated in Fig S3**,** proceeds as follows. First for the strains under study we enumerate all possible unrooted trees. Next, for each phylogenetically informative SNP in the genome, we determine which trees are compatible with and incompatible with the pattern of variation shown at that SNP. In the next step, for each tree we determine all blocks in the genome that are in

agreement with that tree by assigning regions of the genome with consecutive compatible SNPs to single continuous block, and allowing half of the genome between a concordant SNP and a disconcordant SNP to be included in the block. Finally, all trees are examined to determine which has the largest block, this block is assigned to the tree, then the segments in each tree are updated to account for this, and this is repeated until every position in the genome is assigned to a block.

To implement the recursive method on our dataset, we first disregarded the data from the House Finch MG isolates. The House Finch MG isolates have too little genetic variation to usefully determine spatial patterns of recombination and were nearly genetically identical to the TK_2001 poultry isolate. By only using the MG poultry isolates and the reference genome, it is possible to work with the full enumeration of possible unrooted trees as there are only 15 and so we could avoid approximate and heuristic methods. The distribution of sizes of phylogenetically concordant blocks is shown in Fig S4, which also displays a distribution obtained by randomly rearranging the patterns of genetic variation shown at each SNP to different positions in the genome.

Fig S4 shows that the signal of recombination in our dataset is not due to a few rare transfer events, but that these genomes are reasonably mixed, as there are a large number of sizable concordant blocks that are in agreement with different trees. We note that we can also test for recombination by creating permutated datasets that keep the position of SNPs fixed and randomly reassigning the patterns of genetic variation shown at each SNP. If spatial clustering is significant, then the number of blocks required to assign the entire genome to a segment should be much less than the number required in a permutated dataset. Fig S5 shows the distribution of blocks required in 2,600 random permutated datasets, and as expected the total number of blocks required is much greater than that required in the actual dataset, again indicating that recombination is statistically significant with a vanishingly small p-value.

Text S7: Effect of recombination on the estimated substitution rate and demonstration of true temporal signal

The presence of recombination could bias our estimate of the substitution rate as inferred from BEAST. The MCMC algorithm used within BEAST proposes and evaluates the parameters in an evolutionary model based on a single phylogenetic tree. However, in the presence of recombination, there is no single phylogenetic tree that represents the history of all of the genomes sequenced, and this discrepancy between the biological reality and the inference model could affect our results. Although we hope future computing developments that use the ancestral recombination graph approach will eventually solve this problem by allowing the current Bayesian inference approaches to account for recombination, at present there are no available methods to systematically perform simultaneous inference of the posterior distributions for all the evolutionary parameters in circular-genome datasets as large as ours. However, despite this difficulty, it is clear that the single best point estimate for the mutation rate will always be on the order of 10^{-5} per site per year, and that given a number of well supported assumptions, that the interval of uncertainty around this estimate will encapsulate this rate to within an order of magnitude.

To demonstrate that our conclusions are robust to the presence of recombination, we note that a simple method of inference which is less affected by recombination gives virtually identical results. A naïve estimate of the mutation rate can be obtained for any two pair of sequences simply by dividing the number of mutations that appear between the earlier sample and the later sample by the amount of time separating the two samples. This method does not require that no recombination has occurred, however it does require that every element in the present genome has diverged for an equal amount of time from the genome it is being compared to. For example, if two genomes are thought to be diverged by 20 years, but lateral gene transfer (LGT) has introduced into the genome some segments that are diverged by over 40 years, than these segments will bias the mutation rate upwards if the 20 year period is used for the entire genome comparison as these more diverged segments likely contain more mutations. Although this makes LGT events typically problematic for these simple rate estimates, due to the host-shift observed in this system, the assumption of equal divergence times for all segments of the genome that differ between our older and newer samples is likely met. Based on the genetic evidence in this paper, the host-shift appears to be a single founder event that created an isolated population with no additional inputs from the source poultry population. This implies that even if recombination is ongoing between the 1994 and 2007 samples, since all of the strains in the population had a recent common ancestor near 1994, any segments introduced by LGT between 1994 and 2007 should be as diverged as the segments they are replacing.

While we would not expect the value of this estimate to be biased by recombination, this naïve estimate is biased towards a higher rate because simply dividing by the difference between the dates when strains were sampled does not account for the time between the last common ancestor of the two samples and the time of initial sampling, which is additional time during which mutations could appear. However, the nature of our data is such that this bias is very small, and any realistic correction for this bias does not substantially change the inference. The reason for this is that the most common ancestor of all of the House Finch MG strains was almost certainly present near the time of our initial sampling period, as supported by three lines

of evidence. First the epizoodic was very well documented as beginning in 1994 by a wide variety of observers, and despite ample opportunity there were no reports of MG infection in House Finches before this date. Second, and in agreement with 1994 being the first year when MG infected House Finches, in a broad sampling of MG from a variety of host species, all of the House Finch MG strains were genetically identical, despite a large amount of diversity in the poultry population, indicating a recent founder event (Fig S1). Finally, our genome level sequencing of the 1994 strains provides additional evidence for this interpretation. The 1994-1995 samples are characterized almost exclusively by singletons (Table S2), indicating a recent common ancestor and population expansion, and therefore a small bias in the naïve estimate. Therefore, given that there was a bottleneck in the founding of the house finch MG strains, the excess time not accounted for by the difference in sampling times is expected to be very small, on the order of a few months compared to the 13 year interval between the 1994 and 2007 samples, meaning that this naïve method, equivalent to a Poisson regression, will provide a very good estimate of the substitution rate. Evaluating this naïve estimate over any given pairwise comparison of 1994 and 2007 strains we get an estimated rate of 1.35-2.36 X 10^{-5} with an average of 1.7 X 10^{-5}. Although calculating an interval of uncertainty around these estimates is dependent on assumptions about the evolutionary process, one assumption that is uninfluenced by the effects of recombination is to assume that mutations are introduced into the genome as a constant Poisson process. With this assumption, the lower interval for the 95% confidence interval of our mutation rate is still on the order of 10^{-5} for the strains in this study. Although violations of a constant Poisson process are some of the most frequent findings in the field of molecular evolution, correctly identifying and modeling such deviations would require much broader sampling of bacteria then this study, or any other published study we are aware of, could provide. However, all indications are that such violations are not large in magnitude (Fig 4), and even if the width of the 95% confidence interval for the rate estimate assuming a Poisson process is doubled in size, the lower bound of the confidence interval is still approximately 10^{-5}. Therefore, we find no plausible violations of the model large enough to substantially alter our rate estimate more than an order of magnitude.

Additionally, as a simple test and demonstration that our data do contain a true temporal signal and the estimated rate is also not an artifact of the BEAST analysis, we used the program Path-O-Gen to evaluate the clock like nature of the data. An ML tree without an assumed clock was first estimated using the program PhyML [13] and the HKY substitution model used in our BEAST analysis. The regression in Path-O-Gen obtained an estimated rate of 1.45 X 10^{-5} using the default root for the tree (R^2 = .68) and it estimated a rate of 9.6 X 10^{-6} (R^2 =.92) using the best-fitting root, confirming that our Poisson regression results and the BEAST analysis are in agreement with this separate method of estimation. Finally, we also performed a randomization test as described in [14] by randomly reassigning the dates of all of our House Finch strains and rerunning our BEAST analysis We performed this randomization 20 times and each time obtained an HPD interval for the rate that did not overlap with our current estimate and was below our current estimated interval.

Text S8: CRISPR Analysis

To annotate the CRISPR elements in our 454 data and Illumina data we designed the program that computationally reconstructs a CRISPR locus from the sequence data, and simultaneously provides visualization tools that allow the user to validate the computational reconstruction, detect polymorphism, and manually check any ambiguities that may appear during the reconstruction. Each read generated from a sample is checked to see if it contains a sequence similar to the CRISPR repeat present in the ancestral genome. Any reads that contain such a sequence are selected into a subset of reads for further inspection. For each read in this subset, a dynamic alignment algorithm is used to identify the portions of the read that belong to the CRISPR repeats, and by exclusion, those portions that belong to the CRISPR spacers. Spacers that are exactly the same, or that appear to only differ due to sequencing errors, are then identified as spacer families, and these families are each given a numeric name determined by their (essentially random) order of discovery. Finally, as in many modern genome assemblers, the complete CRISPR locus is reconstructed through means of a graph. Each spacer family represents a node that can be placed either upstream or downstream of other families that appear in the same read as themselves. The program constructs this graph, determines the order of the observed spacer families, and plots it in simple format for the user.

We used this method to reconstruct the complete CRISPR locus in all 16 of our samples (Table S12). Of the 61 unique CRISPR spacer regions present in the reference genome, none are present in our samples, which collectively have gained a net total of 47 unique spacers since the time of their divergence from the reference genome. Table S8 shows the number of CRISPR spacers in each strain. Our GA_1995 sample has a copy of every spacer found in every other HF MG strain as well as TK_2001, and thus all other House Finch MG genotypes can be represented by deleting or duplicating the CRISPR spacers found in this strain. As such, the CRISPR array in each strain can be represented as a vector of discrete character states. Because adjacent CRISPR spacers are likely to be lost by the same deletion events, we reconstructed the CRISPR tree (Fig 5) using a parsimony method that always allowed deletions of neighboring CRISPR regions to be scored as single events. Following this assumption, we grouped strains into clades based on the presence of shared deletions or duplications. In instances where two or more equally parsimonious explanations could be provided for a pattern of deletions, we represented this ambiguity in the tree (for example the pattern of deletions shared by 2001 and 2007 can be equally well explained by having or not having these groups share a common ancestor to the exclusion of other strains).

Like the SNP phylogeny, the CRISPR phylogeny is consistent with a single origin of the epizootic and implies periodic replacement of the standing genetic variation in successive cohorts (2001, 2007). Although the deletions and expansions of most of the CRISPR spacers shown are likely due to strand slippage or recombination between the CRISPR repeats, the loss of the CRISPR spacers at the start of the locus in the 2007 strains is part of a much larger deletion of 12.7 kb that is unique to these strains and involves an alternative mechanism (deletion 4 in Table S7).

A recent investigation of CRISPR spacer repeats in *Yersinia pestis* found that a majority of spacers found in the CRISPR array originated from other areas of the organisms genome,

indicating that the CRISPR loci might be involved in regulating intra-genome dynamics, such as controlling gene expression levels or IS/prophage proliferation[15]. To determine if this was also true for the spacers we observed in *Mycoplasma gallisepticum,* we blasted each of the 302 unique spacer sequences that we found against the reference genome (blastn, with parameters "*-W 7, -e 1, -F F -r 2"*), and looked for any sequences that had an alignment score over 40 (equivalent to a ~66% match) to the reference genome. This analysis showed that 3 of the 302 spacer sequences were perfect matches to other portions of the genome, meaning they likely originated from proto-spacers within the MG genome. One spacer found within the reference genome was derived from a sequence within a hypothetical membrane protein (annotated MGA_0908), and another from the reference genome was a perfect match to a segment of DNA topoisomerase IV subunit A (MGA_0056). The third perfect match was from the CK_1996 strain, which contained a spacer sequence derived from a VlhA.4.01 lipoprotein gene (MGAH_0966).

To determine if any of the other CRISPR spacers were similar to any previously sequenced organisms, we blasted each of the 302 spacers against the NCBI 'nr' blast database and examined the top hit after excluding any hits to MG genomes. The top scoring hit only had a score of 50, and the top 5 hits were to *Schistosoma mansonis,* Human*,* and Zebrafish, leading us to conclude that there we no significant matches. Based on these comparisons to known DNA within and outside of the MG genome, we concluded that the source of the CRISPR spacers in this study is predominantly was from previously unstudied organisms.

Fig S1: Broad sampling of House Finch and poultry MG strain diversity.

To understand the broad phylogenetic diversity of House Finch and poultry MG strains, guide our choice of poultry strains for genomic sequencing and compare mutation rates in the HF and poultry MG population, we used DNA sequence data from Ferguson et al. [5] to generate a multisequence alignment for 82 MG strains collected from four host species (Turkey, Chicken, House Finch and Gold Finch). This data, henceforth the **L**arge **S**ample **M**ultiple **S**equence **A**lignment, **LS-MSA)** was composed of four gene fragments (from pvpA, mgc2, gapA and an unnamed surface lipoprotein) that when concatenated yielded approximately 1.9 kb of sequence data per strain (with the exact length of each strain varying due to small indels). We added to this dataset sequences for 8 of the 12 House Finch MG strains sequenced in this study that had complete coverage for these gene fragments. The four strains from this study not incorporated into the dataset (TN_1996, GA_1995, AL_2001_53 and AL_2007_05) were excluded because there was not enough sequencing data to accurately assemble the relevant fragments. We also excluded 3 strains from the original work[5] where we could not identify the host-animal species, leaving 82 strains in the final multiple sequence alignment. In this alignment, all the House Finch haplotypes were identical, except for the 2007 strains that differed from the others at two adjacent nucleotide positions.

Certain sections of the gene fragments in the LS-MSA were polymorphic due to insertions/deletions of tandem repeats, and because there is no clear criteria by which to assign the locations of these repeats in an alignment for phylogenetic purposes, for analysis purposes we reduced the ~1.9kb of sequence down to1,363 bp that could be confidently aligned.

Fig. S1. Phylogenetic tree of 82 avian MG strains inferred from four concatenated gene-segments, totaling 1,363 bp, using Neighbor-joining in PHYLIP. Due to recombination in *Mycoplasma gallisepticum*, this single tree may not be completely representative of the organismal history of the strains from which the gene segments were sampled. However, the pattern showing poultry hosts interspersed amongst the leaves of the tree and high diversity within the MG population is also present in neighbor-joining trees separately inferred for each individual gene fragment, consistent with frequent host-shifts by MG. Strain K4366GF97_10 is from an American Goldfinch (*Carduelis tristis*), also a songbird and the chicken reference strain used to obtain the reference genome is R63_44.

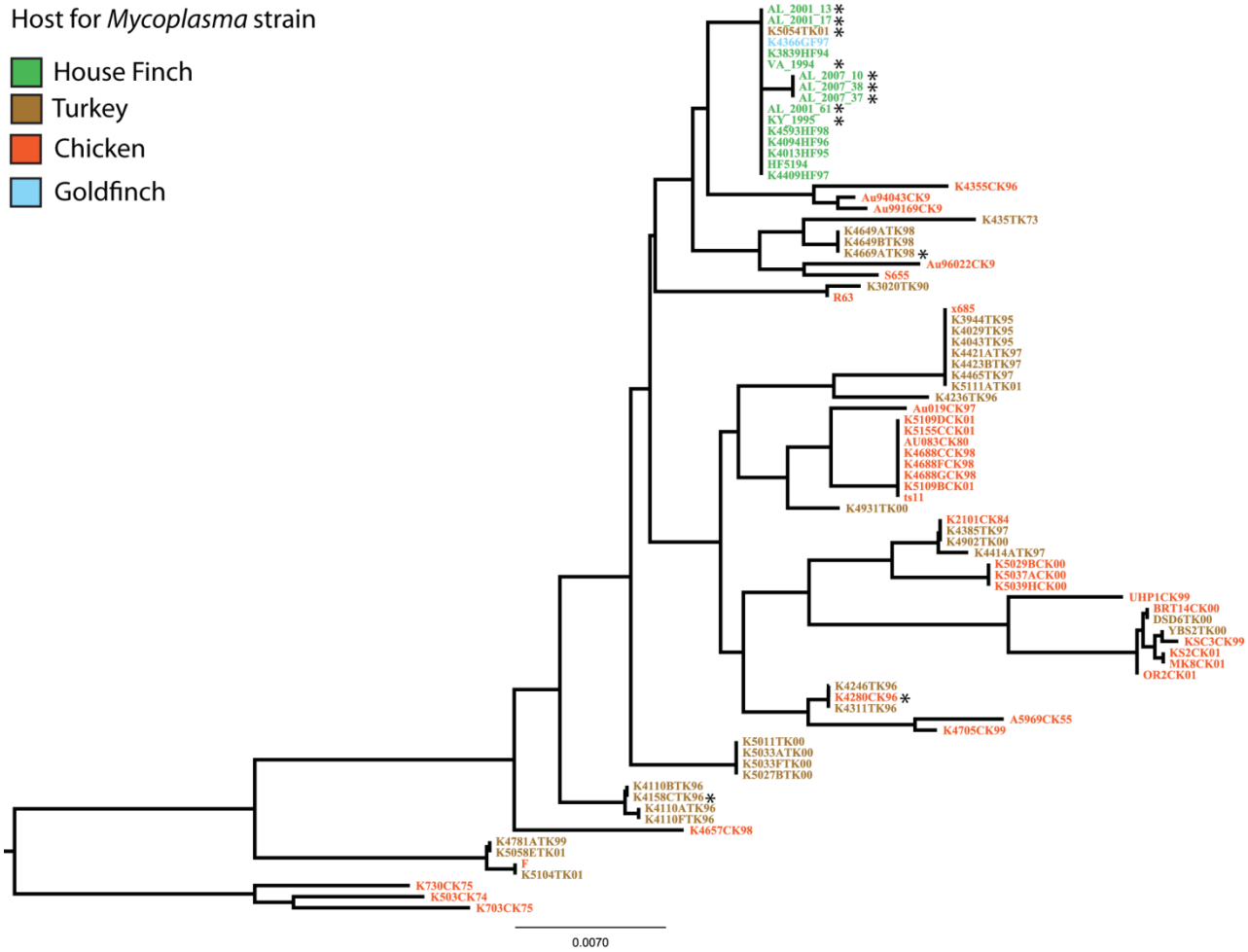


Fig S2: Cross Validation of the 454 Sequencing Data with the Illumina Sequencing Data

Our dataset provides an opportunity to validate the SNP calls made with our 4X-19X coverage 454 data for the House Finch MG isolates by using the SNP calls made with the 294X coverage Illumina data that was generated for TK_2001. TK_2001 and the House Finch MG isolates (particularly the pre-2001 isolates) are nearly genetically identical, and SNPs for both strains were called relative to the much more distantly related strain that was used to generate the reference genome. As outlined with the unrooted tree shown in Fig S1 this means that most of the SNPs called for each of the House Finch isolates should also be called for the TK_2001 strain, with any unmatched SNPs likely due to either genetic divergence between the two strains or SNP calling errors.

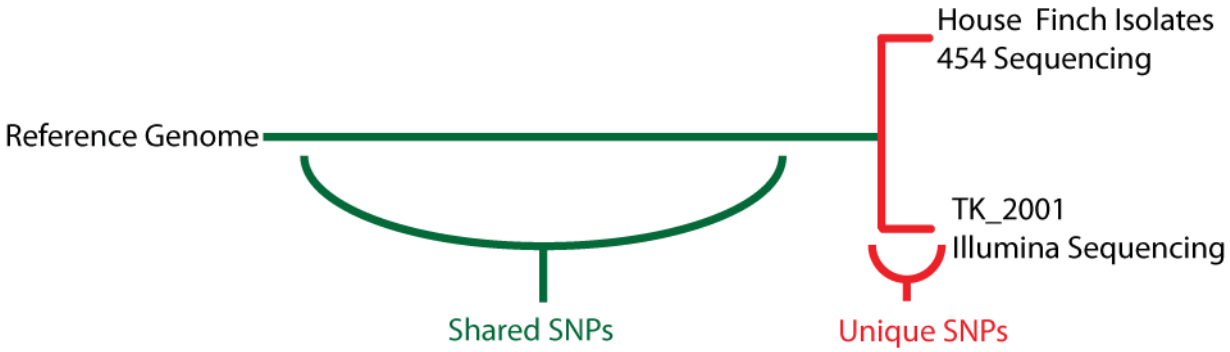


Fig. S2 Schematic representation of the cross validation approach.

The House Finch isolates that were sequenced with the lower coverage 454 data are very closely related to the TK_2001 strain sequenced with Illumina technology. Since SNPs in both TK_2001 and the House Finch isolates were called against the reference genome, comparing the high coverage TK_2001 Illumina strains to the House Finch Isolates allows us to estimate the total number of SNPs that could either be due to false positives or genetic divergence between the strains.

The results of this comparison are shown in table S4**.** For our most stringent threshold, of the up to 6,461 SNPs that were called in our pre-2001 House Finch isolates, 99.7% of the SNPs called with the 454 data were also called with the Illumina data. This bounds the false positive rate for SNP calls in the 454 stringent data at 0.3%. However, we believe that this unmatched 0.3% is due to true genetic divergence between the strains and not sequencing errors, as these SNPs are very well supported. For example, all 21 SNPs in VA_1994 that did not match TK_2001 were supported by at least 9 reads that contained the variant, and often many more.

Table S4 documents the robustness of our population genetic estimates on variations in SNP calling protocol, leading only to minor variations (~1%) in the false positive rate for our SNP datasets. This shows that almost all of the uncertainty in estimating the mutation rate from these genomes is due to the inherent sampling variability that naturally results from the stochastic process that generated them and is not due to any variability that comes from calling SNPs in

these genomes. Additionally the ratio of polymorphic to conserved sites is equivalent across all three datasets.

Fig. S3. Illustration of the recursive method used to assign segments of the genome to phylogenetically concordant blocks.

At the initialization of the algorithm the phylogenetically informative SNPs in the genome (x's in the diagram) are used to determine continuous segments that are in agreement with all possible trees. Sections of a genome in agreement with a particular tree are shown as solid colored lines over that genome segment. Note that any one SNP can be in agreement with multiple trees. If only one of two adjacent SNPs are in agreement with a tree, then half of the distance between the two SNPs is assigned to the concordant segment.

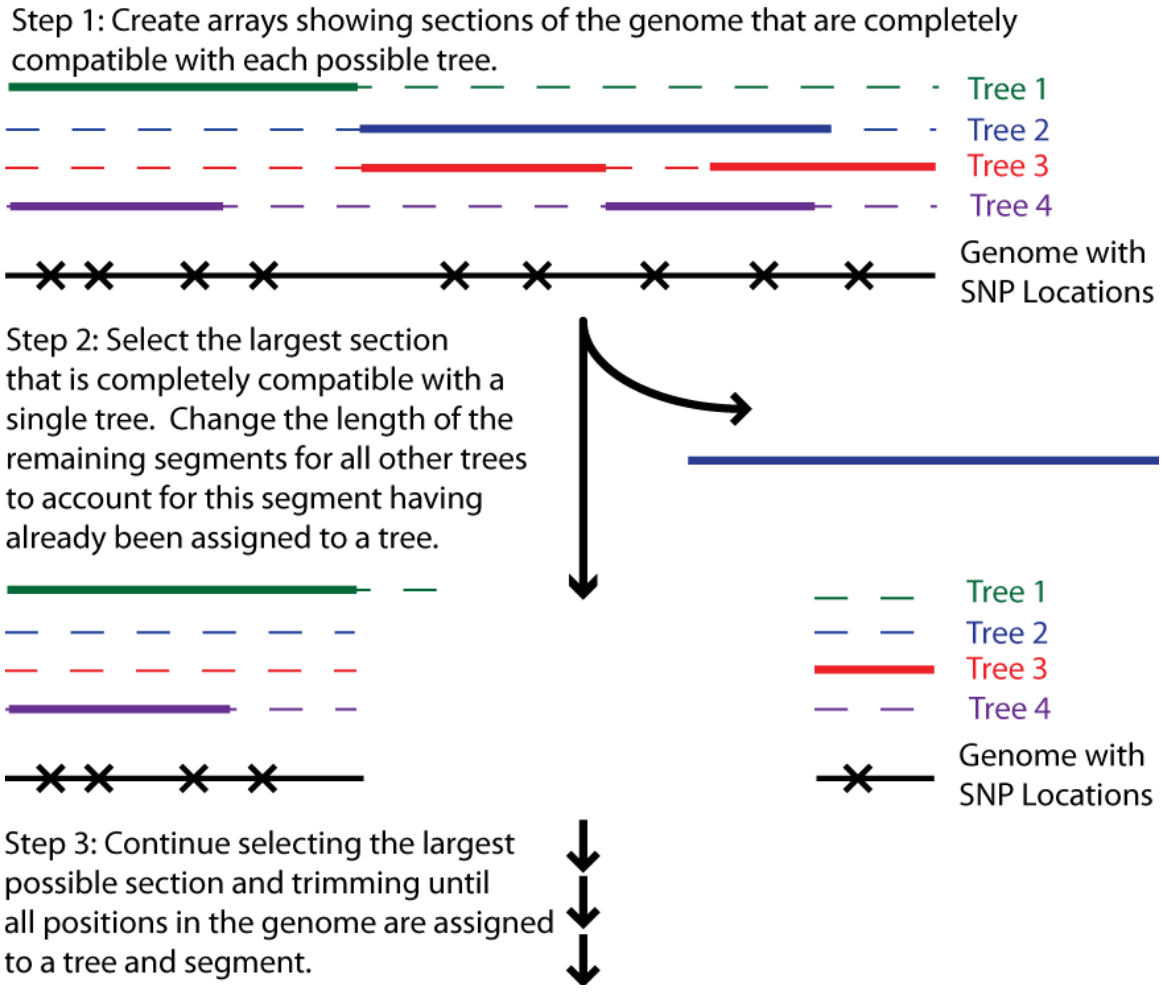


Fig. S4. Distribution of the number of phylogenetically concordant segments in the genome and in a dataset obtained by a single random permutation of the SNPs. Block sizes are in bp.

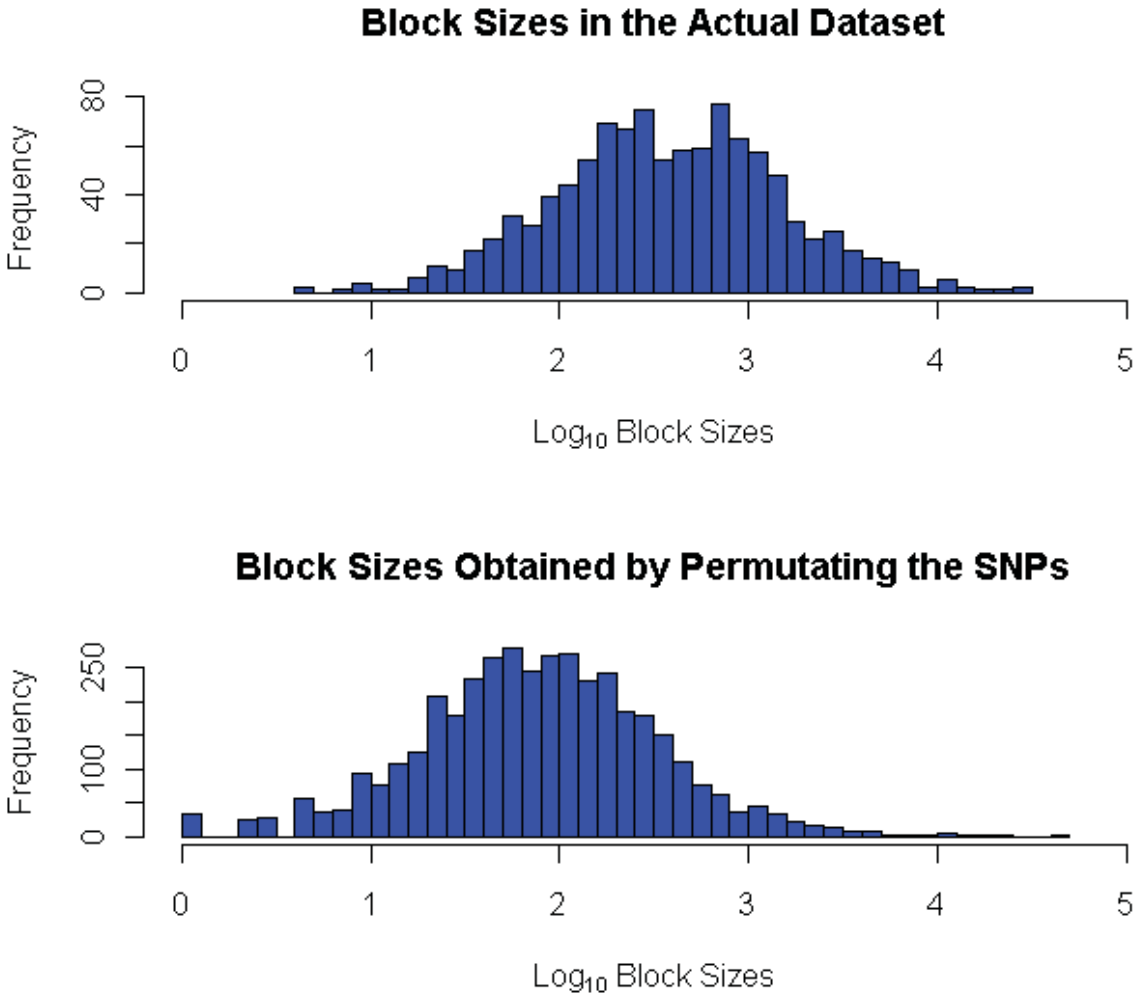


Fig. S5. Distribution of the size of phylogenetically concordant segments in the genome and in a dataset obtained by repeatedly creating permutations of the SNPs.

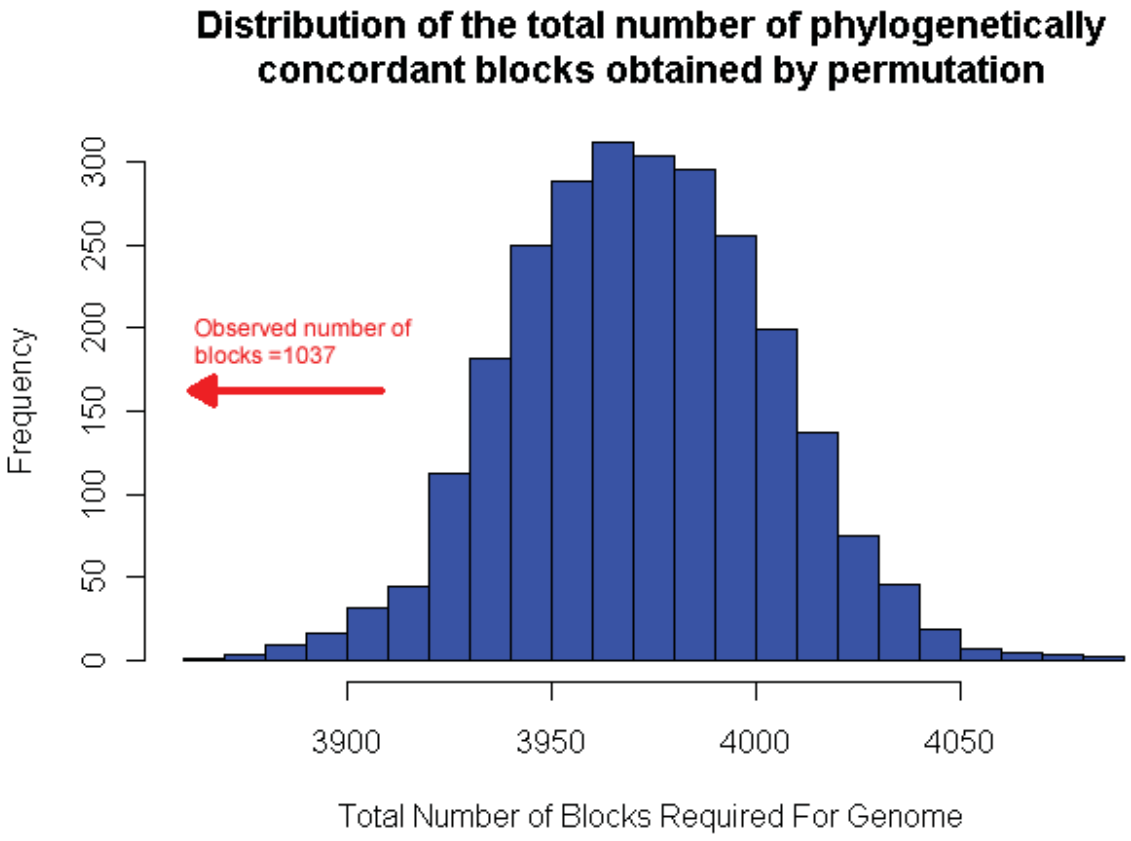


Fig. S6. 95 % HPD intervals of the rate estimated in BEAST using our actual dataset, as well as 20 permutations of the data where the dates on the tips are randomly reassigned. The interval for the true dataset is shown in red, and the randomized datasets are shown in blue.

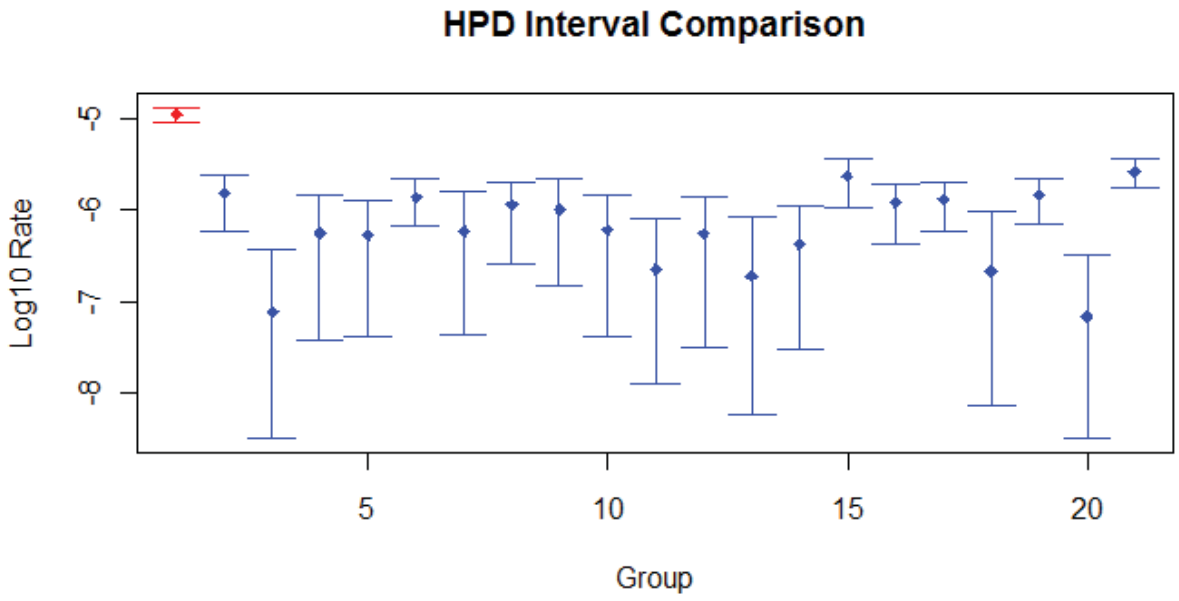


Table S1: Isolates used

We studied 12 field isolates of *Mycoplasma* collected from House Finches in the Southeastern United States. The isolates were chosen to encompass the complete time span of the epizootic with four samples from the 1994-1996 period, four from 2001 and four from 2007. We also studied four isolates of *Mycoplasma* collected from poultry. Table S1 below shows the source of the isolates, their sequencing coverage in terms of the reference genome[1], and any alternate names the strains may have had in previous studies.

Table S1. Characteristics of MG isolates used in study

Strain Name	Host species*	Coverage	Avg. Quality Score	Date Isolated	Isolated From	Source	Alternate Name
AL_2001_13	HF	11.4	27	March 6, 2001	Lee County, Alabama	This study	
AL_2001_17	HF	8.9	18	June 27, 2001	Lee County, Alabama	This study	
AL_2001_53	HF	6.5	24	March 14, 2001	Lee County, Alabama	This study	
AL_2001_61	HF	9.5	23	February 11, 2001	Lee County, Alabama	This study	
AL_2007_05	HF	8.4	34	January 20, 2007	Lee County, Alabama	This study	
AL_2007_10	HF	4.3	37	January 20, 2007	Lee County, Alabama	This study	
AL_2007_37	HF	18.9	23	February 11, 2007	Lee County, Alabama	This study	
AL_2007_38	HF	9.8	35	February 11, 2007	Lee County, Alabama	This study	
GA_1995	HF	7.2	27	February 13, 1995	Clarke County, Georgia	[17]	K3891
KY_1996	HF	7.3	22	February 26, 1996	Kentucky	[18]	K4117
TN_1996	HF	6.8	24	January 23, 1996	Shelby County, TN	[18]	K4094
VA_1994	HF	13.9	24	June, 1994	Virginia	[19]	S11
TK_2001	Turkey	294	33.4	2001	Indiana	[5]	K5054TK01
TK_1998	Turkey	391	33.4	1998	Colorado	[5]	K4669ATK98
TK_1996	Turkey	498	33.4	1996	Missouri	[5]	K4158CTK96
CK_1996	Chicken	460	33.4	1996	Missouri	[5]	K4280CK96

*HF = House Finch

The House Finch isolates from 2001 and 2007 were obtained for this study as follows. House finches were caught in wire mesh cages placed around feeders and in mist nets. Upon capture, *Mycoplasma* samples were collected by swabbing eye conjunctiva and choanal cleft of birds displaying symptoms of disease. Swabs were immediately placed into 3 mL of SP4 media preheated to 37 C. After gentle vortexing, the swab was removed and the inoculated broth [20] was incubated at 37 C overnight. After approximately 24 hours, a 1:10 blind passage was performed for each culture which was then incubated at 37 C for 5 weeks or until a color change indicated growth [21]. Following a media color change, stocks of each isolate were made as follows: 500uL of a 1:1 solution of SP4 broth and glycerol was added to 500uL of cell culture. Samples were gently mixed and frozen at -80 C for long-term storage. DNA for sequencing was prepared by re-inoculating frozen cultures into SP4 media incubated until log phase. DNA was extracted from each sample at between passage five and seven using Qiagen DNA tissue minipreps.

Table S2. SNP counts in the final working data set comprising the 17-way alignment

	All	House Finch Strains	House Finch Strains and TK_2001	1994-1996 Strains	2001 Strains	2001 Strains excluding AL_2001_17	2007 Strains	New Poultry Strains and Reference Genome excluding TK_2001	New Poultry Strains
Total SNPs	16,398	412	469	136	152	42	37	14,400	13,175
Synonymous	9,383	122	138	37	50	12	11	8,459	7,735
Non-synonymous	5,324	246	279	85	88	24	21	4,534	4,090
Non-coding	1,729	45	53	14	15	7	5	1,441	1,377
Singletons	8,576	258	310	115	103	42	36	8,208	5,601
Phylogenetically informative within the group	7,693	152	157	20	48	0	0	6,048	7,517
Fixed SNPS (Ignoring missing data)	N/A	80	1,579	1	3	29	87	1,551	140
Fixed SNPS (Require data from all group members)	N/A	8	310	0	0	20	47	1,459	55
Fixed SNPS (Require data from all strains in study)	N/A	0	301	0	0	9	24	297	0
Fixed SNPS (Require data from all non-group members, but allows incomplete data within specified group)	N/A	2	1,485	0	0	12	36	306	0

Table S3. Sanger Sequencing Validation of SNP Calls

In the early stages of the project we validated a small subset (n = 9) of SNPs via PCR and direct sequencing of 76 sequencing reactions spread across the 12 House Finch strains. We selected these sequenced positions for two reasons. First, these sites were phylogenetically informative for the pre-2001 House Finch strains whose relationships we wished to resolve. Second, we felt these SNPs were the most suspect of all of the SNPs in our dataset as they provided conflicting phylogenetic information and so were either strong evidence for an unknown source of sequencing errors in our methods or strong evidence for recombination in this population of *Mycoplasma*. We were able to rule out sequencing error as all sequenced loci confirmed the polymorphisms identified by the 454 sequencing (71 of 76 loci matched the 454 sequencing data, and 5 of 76 provided data for strains that did not have adequate coverage at that position in the original 454 data).

Table S3. SNPs Validated by PCR amplification and Sanger Sequencing

Strain	Position	Sanger bp	Reference bp	454 bp
AL_2001_13	170360	C	C	C
AL_2001_17	170360	T	C	T
AL_2001_61	170360	C	C	C
AL_2007_05	170360	T	C	T
AL_2007_10	170360	T	C	T
AL_2007_37	170360	T	C	T
AL_2007_38	170360	T	C	T
GA_1995	170360	C	C	C
KY_1996	170360	C	C	C
TN_1996	170360	C	C	N
VA_1994	170360	C	C	C
AL_2001_13	174643	C	C	C
AL_2001_17	174643	C	C	C
AL_2001_53	174643	C	C	C
AL_2001_61	174643	C	C	C
AL_2007_05	174643	T	C	T
AL_2007_10	174643	T	C	T
AL_2007_37	174643	T	C	T
AL_2007_38	174643	T	C	T
GA_1995	174643	C	C	C
KY_1996	174643	T	C	T
TN_1996	174643	C	C	C
VA_1994	174643	C	C	C
AL_2001_13	580857	G	G	G
AL_2001_61	580857	G	G	G
AL_2007_05	580857	T	G	T
AL_2007_37	580857	T	G	T
AL_2001_13	691180	G	G	G

Table S3 (continued).

AL_2001_17	691180	A	G	A
AL_2001_61	691180	G	G	G
AL_2001_61	691180	G	G	G
AL_2007_05	691180	A	G	N
AL_2007_10	691180	A	G	A
AL_2007_37	691180	A	G	A
AL_2007_38	691180	A	G	A
GA_1995	691180	G	G	G
KY_1996	691180	G	G	G
TN_1996	691180	G	G	N
AL_2001_17	716811	C	C	C
AL_2007_05	716811	C	C	C
AL_2007_37	716811	C	C	C
AL_2007_38	716811	C	C	C
TN_1996	716811	C	C	C
VA_1994	716811	C	C	C
AL_2001_13	720901	T	T	T
AL_2001_17	720901	T	T	T
AL_2001_53	720901	T	T	T
AL_2001_61	720901	T	T	T
GA_1995	720901	T	T	T
TN_1996	720901	T	T	T
AL_2001_13	853947	A	G	A
AL_2001_17	853947	G	G	G
AL_2001_53	853947	A	G	A
AL_2001_61	853947	A	G	A
AL_2007_05	853947	G	G	G
AL_2007_10	853947	G	G	N
AL_2007_37	853947	G	G	G
AL_2007_38	853947	G	G	G
GA_1995	853947	G	G	G
KY_1996	853947	A	G	A
TN_1996	853947	G	G	G
VA_1994	853947	A	G	A
AL_2001_13	909457	A	C	A
AL_2001_61	909457	A	C	A
AL_2007_05	909457	C	C	C
AL_2007_37	909457	C	C	C
AL_2007_38	909457	C	C	C
VA_1994	909457	C	C	C
AL_2001_13	973203	G	G	G
AL_2001_17	973203	G	G	G
AL_2001_53	973203	G	G	N
AL_2001_61	973203	G	G	G
AL_2007_37	973203	A	G	A
AL_2007_38	973203	A	G	A
GA_1995	973203	G	G	G
VA_1994	973203	G	G	G

Table S4. Cross validation of the 454 SNP calls using the Illumina SNP calls

Alignment File Name	**Stringent_2010_Masked_4_Val.fna**											
Strain	**TN_1996**	**GA_1995**	**KY_1996**	**VA_1994**	**AL_2001_53**	**AL_2001_17**	**AL_2001_61**	**AL_2001_13**	**AL_2007_10**	**AL_2007_05**	**AL_2007_38**	**AL_2007_37**
Differences from Reference	3129	2613	5206	6482	4044	5682	5327	5770	3772	3260	5766	6732
Differences Shared with TK_2001	3121	2604	5194	6460	4022	5621	5292	5737	3723	3212	5694	6642
% Identical SNP calls	99.7%	99.7%	99.8%	99.7%	99.5%	98.9%	99.3%	99.4%	98.7%	98.5%	98.8%	98.7%
Singletons for Strain	6	7	6	15	3	54	9	10	3	3	2	7

Alignment File Name	**Stringent_Moderate_v2_2010_Masked_4_Val.fna**											
Strain	**TN_1996**	**GA_1995**	**KY_1996**	**VA_1994**	**AL_2001_53**	**AL_2001_17**	**AL_2001_61**	**AL_2001_13**	**AL_2007_10**	**AL_2007_05**	**AL_2007_38**	**AL_2007_37**
Differences from Reference	5402	4763	7012	7482	6118	7269	7120	7347	5722	5379	7181	7428
Differences Shared with TK_2001	5352	4690	6972	7437	6045	7173	7053	7282	5628	5275	7067	7307
% Identical SNP calls	99.1%	98.5%	99.4%	99.4%	98.8%	98.7%	99.1%	99.1%	98.4%	98.1%	98.4%	98.4%
Singletons for Strain	26	58	15	17	29	65	9	5	9	21	7	4

Alignment File Name	**Moderate_2010_Masked_4_Val.fna**											
Strain	**TN_1996**	**GA_1995**	**KY_1996**	**VA_1994**	**AL_2001_53**	**AL_2001_17**	**AL_2001_61**	**AL_2001_13**	**AL_2007_10**	**AL_2007_05**	**AL_2007_38**	**AL_2007_37**
Differences from Reference	6411	5875	7336	7682	6719	7526	7487	7638	6191	6380	7439	7598
Differences Shared with TK_2001	6306	5699	7262	7615	6553	7399	7374	7545	6017	6186	7291	7456
% Identical SNP calls	98.4%	97.0%	99.0%	99.1%	97.5%	98.3%	98.5%	98.8%	97.2%	97.0%	98.0%	98.1%
Singletons for Strain	70	151	36	25	108	81	38	16	77	78	16	6

Table S5. Estimates of genetic diversity (π) in subgroups of MG strains sampled from different host species* in the LS-MSA

Host species, year	N	bp	π	Standard Deviation
All	73	~1362	0.01963	0.00106
Chicken, all	26	~1362	0.01888	0.00171
Chicken, 1994-1996, inclusive	4	~1362	0.01853	0.00397
Chicken, 1994-1996 (no Australia samples)	2	~1362	0.02428	0.01214
Chicken, post-1996	18	~1362	0.01737	0.00191
All turkey	31	~1362	0.02253	0.00193
Turkey, all	33	~1362	0.02203	0.00159
Turkey,1994-1996, inclusive	10	~1362	0.01634	0.00161
Turkey, post-1996	21	~1362	0.02332	0.00201
House finch, all	14	~1362	0.00057	0.00019
House Finch, this study	**12**	**743,011**	**0.00014**	**0.00001**
1994-1996	**4**	**743,011**	**0.00010**	**0.00003**
2001	**4**	**743,011**	**0.00011**	**0.00004**
2007	**4**	**743,011**	**0.00003**	**0.00001**

Data from this study (bold) and from Ferguson et al. 2005 (5).

Table S6: Patterns of synonymous and nonsynonymous substitutions

We compared the frequencies of non-synonymous, synonymous and non-protein coding SNPs in the House Finch and poultry populations by comparing three groups of SNPs. The first type were polymorphisms that likely arose in the House Finch MG lineage, as they are fixed in the poultry MG strains but are polymorphic amongst the House Finch ones. The second group are those SNPs that likely arose in the poultry MG population, as they show the opposite pattern and are fixed in the House Finch strains. We also examined SNPs that represented fixed differences between the two populations and likely arose on the lineage separating the poultry and House Finch MG populations. 35 SNPs were excluded from categorization because they were polymorphic in both the poultry and House Finch populations. Finally, we obtained expected numbers of the three types of mutations by simulating mutations in the genome using the maximum a posteriori parameters for the HKY substitution model inferred from our earlier BEAST analysis (Text S2).

Observed and expected number of SNPs in various comparisons among strains.

	Polymorphic within House Finch strains	Polymorphic within poultry strains	Fixed Differences	Simulated
Synonymous	28.5%	58.0%	27.8%	25.7%
Nonsynonymous	59.9%	31.6%	40.5%	63.8%
Non-protein Coding	11.6%	10.4%	31.6%	10.5%
Total SNPs	379	15,940	79	10,000

By converting table S6 into a contingency table, one can reject the assumption that the mutations are distributed as one would expect under neutrality as defined by the simulated distribution in the poultry population, but not in the House Finch population. ($p_{poultry}$<2.2e-16, p_{HF}=.28), which is consistent with other studies that have shown very recently diverged pathogens tend to evolve neutrally [22].

To obtain estimates of the distribution of dN/dS values for each gene within MG from all of our samples using PAML v. 4.2b[23]. For each gene, we used the maximum clade credibility tree from our BEAST analysis (Test S2) and for those genes that contained both non-synonymous and synonymous mutations we used PAML to estimate the dn/ds (omega) ratio. These data are summarized in Fig. 2c of the main text.

Table S7. Regions of the reference genome that had been lost in House Finch MG isolates

We searched for genes in the reference genome that were not present in the House Finch MG isolates. The 454 contigs assembled from our pooled 2007 samples were mapped to the reference genome using Megablast and any portion that aligned with greater than 95% similarity and over 100 bp in length to a section of the reference genome was considered to represent that section. We then searched for any section of the reference genome longer than 200 bp in length that was not represented by some of the reads in our sample. Unrepresented segments were then further investigated to confirm the deletion, determine the likely mechanism by which it occurred and the starting and ending points in the coordinates given by the reference genome. For the reasons given previously, any putative deletions that appeared in the VlhA regions were not investigated further in this study, even though these regions likely experienced deletions relative to the reference MG strain.

The list of reconstructed deletions in House Finch MG isolates from this analysis is shown in the Table S7; in total they account for ~42 kb of the reference genome being lost and are responsible for the deletion of a total of 34 genes. Three of these deletions are hypothesized to have occurred via recombination between IS elements. Two of the large deletions (numbers 3 and 5) could clearly be identified because no reads representing the deleted sequence were present and because a contig could be formed that spanned the deletion. However, three of the deletions (numbers 1, 3 and 5) were clearly mediated by an IS element insertion followed by a non-homologous recombination-mediated deletion. As these events are caused by recombination between non-homologous sequences, the exact location of the recombination point is unknown and only approximate coordinates are given in the Table S7. All of the deletions found were present in all of our other HF strains, except for the 12.7 kb deletion which was unique to the 2007 isolates.

Deletion number	Approx. start	Approx. end	Deletion size (bp)	Deletion mediated by recombination between IS elements?	Distribution
1	124,815	126,674	1,859	Yes	All strains except R_{low}
2	137,173	138,833	1,660	No	All strains except R_{low} and CK_1996
3	369,420	388,013	18,593	Yes	All strains except R_{low} and TK_1996
4	912,433	925,150	12,717	No	Only in 2007 House Finch strains
5	938,560	945,976	7,416	Yes	All strains except R_{low}
Total deleted: ~42,245 bp					

Table S8. Descriptions of six novel insertion sites of IS elements and insert characteristics for House Finch MG strains.

	Approximate Location	Sides Present	Target Gene	Description of Insertion Area
A	124818	5’	MGA_0801	Potential C-terminal fragment of subtilisin like protease
B	295023	Both	None	This section of the genome is unannotated. The location is 1,047 and 276 bp away from the genes on either side.
C	464795	Both	MGA_1220	ArcA, a predicted arginine deiminase
D	537089	Both	None	This landed inside a pseudo-gene that formerly was an acetyl-CoA hydrolase/transferase
E	560163	Both	None	This is 201 bp and 167 bp away from the nearest genes on either side.
F	938560	5’	None	This is 142 bp and 151 bp away from the genes on either side of this insertion.

Table S9. Genes pseudogenized or deleted in the House Finch MG isolates and their status in other *Mycoplasma* genomes.

Among the 12 House Finch isolates we identified 34 genes that had been removed by a deletion, 2 that had been disrupted by a transposon insertion (including one that was deleted following this insertion) and 17 genes that had been pseudogenized by frameshift or nonsense mutations, for a total of 52 genes . We sought to evaluate if these genes were unique to the reference MG genome by evaluating if they had any homologues in any of the 20 Mollicute genomes available has determined by the Molligen Database [24]. We found that 5 of the 33 genes (15%) lost by a deletion lacked a homologue in at least one other genome, while 3 of the 17 genes lost by pseudogenization in the House Finch isolates (~18%) lacked a homologue in the other genomes. We also checked whether any of the genes that were lost in the House Finch isolates had homologues in every one of the 13 *Mycoplasma* genomes available in the database, and thus could be considered "core" genes. We found that of the 229 genes in the reference genome that had a homologue in all of the other genomes, 7 of these had been lost by a combination of 1 deletion and 3 frameshift mutations in the House Finch MG strains.

Table S9. Genes pseudogenized or deleted in the House Finch MG isolates and their status in other *Mycoplasma* genomes.

Gene ID	Start	End	How Lost	No Homology to other *Mycoplasma* genomes	Homology to all other *Mycoplasma* genomes	Gene Name	Product
MGA_0625a	5159	6077	Disruptor Mutation	FALSE	FALSE		ABC-type multidrug/protein/lipid (MdlB-like) transport system component domain protein
MGA_0626	6392	8294	Disruptor Mutation	FALSE	TRUE		ABC-type multidrug/protein/lipid (MdlB-like) transport system component
MGA_0641	14480	15212	Disruptor Mutation	FALSE	FALSE	glpF	glycerol uptake facilitator protein GlpF
MGA_0648	18209	20462	Disruptor Mutation	FALSE	FALSE		conserved lipoprotein
MGA_0656	30264	30948	Disruptor Mutation	TRUE	FALSE		unique hypothetical lipoprotein
MGA_0686	50592	52587	Disruptor Mutation	FALSE	TRUE	uvrB	excinuclease ABC subunit B
MGA_0687	52631	53789	Disruptor Mutation	FALSE	FALSE	pstS	ABC-type phosphate transport system periplasmic phosphate binding protein
MGA_0801	124459	125605	Deletion/IS Insertion	FALSE	FALSE		Subtilisin-like serine protease domain protein
MGA_0802	125682	126432	Deletion	TRUE	FALSE		Subtilisin-like serine protease domain protein
MGA_0815	137104	139078	Deletion	FALSE	FALSE		Subtilisin-like serine protease
MGA_1037	332335	334084	Disruptor Mutation	FALSE	FALSE		conserved hypothetical membrane protein
MGA_1328	369554	369794	Deletion	FALSE	TRUE	deoC_1	Deoxyribose-phosphate aldolase domain protein
MGA_1081	369839	371102	Deletion	FALSE	FALSE		putative transposase

Table S9 (Continued).

MGA_1083	371070	371919	Deletion	FALSE	FALSE		HAD superfamily hydrolase Cof
MGA_1085	371929	373567	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_1087	373576	374212	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_1088	374241	374925	Deletion	FALSE	TRUE		ABC transporter ATPase component
MGA_1089	374908	376459	Deletion	FALSE	FALSE		ABC transporter permease domain protein
MGA_1091	376555	376891	Deletion	FALSE	FALSE		putative signal peptidase I
MGA_1092	376969	377530	Deletion	FALSE	TRUE		Elongation factor G domain protein
MGA_1100	379435	380476	Deletion	FALSE	TRUE	asnS_2	Asparaginyl-tRNA synthetase
MGA_1102	380479	382111	Deletion	FALSE	FALSE		conserved hypothetical membrane protein
MGA_1103	382094	384026	Deletion	FALSE	FALSE		predicted integral membrane methylase-domain protein
MGA_1347	384502	384670	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1106	384754	384946	Deletion	TRUE	FALSE		putative transposase domain protein
MGA_1107	384995	386495	Deletion	FALSE	FALSE		conserved hypothetical RmuC-domain protein
MGA_1108	386618	387119	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1109	387260	388307	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1220	464277	465489	IS Insertion	FALSE	FALSE	arcA_1	Arginine deiminase
MGA_1263	507145	507823	Disruptor Mutation	FALSE	FALSE	beta-pgm	putative beta-phosphoglucomutase (beta-PGM)
MGA_1283	520043	520808	Disruptor Mutation	FALSE	FALSE		PTS system mannitol-specific (MtlA)-like IIB domain protein
MGA_1305	536425	536824	Disruptor Mutation	FALSE	FALSE	maoC	MaoC-like dehydratase
MGA_0135	652030	653536	Disruptor Mutation	FALSE	TRUE	potA	ABC-type spermidine/putrescine import ATP-binding protein potA
MGA_0137	653891	655376	Disruptor Mutation	TRUE	FALSE		unique hypothetical protein
MGA_1361	747656	747986	Disruptor Mutation	FALSE	FALSE		unique hypothetical protein
MGA_1354	876961	877114	Disruptor Mutation	FALSE	FALSE		hypothetical protein
MGA_0508	910763	912797	Deletion	FALSE	FALSE	fruA	PTS system fructose-specific enzyme EIIABC component

Table S9 (Continued).

MGA_0512	912799	913246	Deletion	TRUE	FALSE		hypothetical protein
MGA_0514	913193	914144	Deletion	FALSE	FALSE	manA	mannose-6-phosphate isomerase (phosphomannose isomerase)
MGA_0516	915226	916669	Deletion	TRUE	FALSE		unique hypothetical protein
MGA_0517	916577	917954	Deletion	FALSE	FALSE		Subtilisin-like serine protease domain protein
MGA_0518	917874	918705	Deletion	TRUE	FALSE		Subtilisin-like serine protease domain protein
MGA_0519	919247	923060	Deletion	FALSE	FALSE		Csn1 family CRISPR-associated protein
MGA_0523	923127	924054	Deletion	FALSE	FALSE	cas1	CRISPR-associated protein Cas1
MGA_0525	924040	924370	Deletion	FALSE	FALSE	cas2	CRISPR-associated protein Cas2
MGA_0526	924369	925134	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_0537	938710	941338	Deletion	FALSE	FALSE	hsdM	type I restriction-modification system methyltransferase (M) subunit
MGA_0539	941547	942165	Deletion	FALSE	FALSE	hsdS_1	type I restriction-modification system specificity (S) subunit domain protein
MGA_0540	942139	942724	Deletion	FALSE	FALSE	hsdS_2	type I restriction-modification system specificity (S) subunit domain protein
MGA_0541	942734	945890	Deletion	FALSE	FALSE	hsdR	type I site-specific restriction-modification system restriction (R) subunit (deoxyribonuclease)
MGA_0567	970243	970549	Disruptor Mutation	TRUE	FALSE		unique hypothetical protein
MGA_0586	987980	990098	Disruptor Mutation	FALSE	FALSE		conserved hypothetical protein

Table S10: Mutations in the *UvrB* Gene and Possible Effects

The *UvrB* gene in every house finch MG strain sampled contains a mutation that truncates the final 3 amino acids of the protein, and this mutation is also present in the closely related TK_2001. The DNA encoding the C-terminal of this amino acid contains a 2 time repeat of the sequence "TAAG" and this mutation introduced one additional repeat of this sequence as a 4 bp insertion. The effect of this 4 bp insertion was to introduce an early "TAA" stop codon and thereby truncate the protein by 3 amino acids as shown below.

Comparison of the C-Terminals in the *UvrB* gene

House Finch MG Isolates	…KMIEDLRNEMLEAAKNQNYEHAASLRDLIIELETQQLSK*
Reference MG Genome	…KMIEDLRNEMLEAAKNQNYEHAASLRDLIIELETQQLSKTNK*

UvrB is an integral part of the cell's DNA excision repair system and functions by forming associations with *UvrA* and *UvrC* during the repair process. Experimental work with the *UvrB* protein from E. coli has shown that the C-terminal of this protein is essential for the protein to associate with *UvrC* and allow a repair to occur [25,26]. However, the house finch MG protein has lost only the final 3 amino acids, and so the specific effect of this mutation cannot be determined from past functional or comparative work.

DNA excision repair is responsible for the repair of pyrimidine dimers, and one signature that these types of mutations have not been repaired along an evolving lineage is the presence of "CC" to "TT" mutations (or "GG" to "AA" if the effect of the mutation is viewed from the other strand). To investigate if the rate of these mutations is elevated in the house finch MG samples, we compared the characteristics of adjacent SNPs that are found segregating amongst the house finch and TK_2001 MG samples to those adjacent SNPs that are polymorphic amongst the reference genome and the other poultry strains. This comparison is shown in table S14.

This comparison showed many features that suggested inhibition of the nucleotide excision repair system within the house MG. The majority of the double mutations within the house finch MG could be identified as involving a "CC" to "TT" substitution on one of the strands of DNA. Among the house finch MG samples, 14 pairs of SNPs were adjacent to each other (Table S11). Of these, 13 could be parsimoniously identified as having occurred on a single, and the same, branch of the tree, and 12 of these could be defined (using the reference and poultry strains to identify the derived allele) as a "CC" to "TT" substitution. Of the two remaining adjacent SNP pairs (at reference positions 667,905 and 715,595), one involved two mutations that occurred on separate branches on the tree, such that no genotype contained a copy of both derived alleles, and another involved an "AA" to "TT" transition. Also suggestive of an increase in the mutation rate for paired bases is the high number of adjacent SNPs given the small number of total SNPs within the HF samples. The percentage of SNPs that are adjacent to each other is expected to increase with the total number of SNPs in an alignment. However, despite having a much smaller number of SNPs, those that were polymorphic among the house finch MG strains contained a greater proportion of adjacent SNPs (Table S11).

We tested for an increase in the number of paired substitutions that involved a substitution from two identical bases to two identical bases of a different type. A contingency table for this analysis was constructed by counting only the adjacent SNPs that appeared in pairs (excluding SNPs that appeared in adjacent groups of three or more, as well as SNPs with over 2 types segregating). The frequency of identical conversions in each group was then compared and found to be significantly different ($p < 0.00001$). This analysis is slightly complicated because at one of the positions containing adjacent SNPs in the house finch MG samples, position 667,905 in the reference genome coordinates, the ancestral sequence "CC" sequence has mutated in one strain to create a "CC"-> "CA" substitution, while on the branch leading to the 2007 strains it has mutated to create a "CC"-> "TT" substitution. Although this made the classification of this pair ambiguous, the frequency difference for these types of mutations suggests that either classification still results in a significant difference, though for clarity we presented it as an identical pair substitution in table S14.

Table S10 – Comparison of adjacent SNPs within the house finch MG to those between the house finch MG and the reference genome.

	SNPs polymorphic amongst strains without the *UvrB* mutation but fixed amongst strains that have it	**SNPs that are polymorphic amongst the strains with the *UvrB* mutation.**
Total SNPs	16,959	420*
SNPs adjacent to another SNP (percentage of total SNPs)	1,458 (8.5%)	28 (6.8%)
Adjacent Pairs of SNPs (excluding >3 SNPs in a row)	641	14
Adjacent pairs with a conversion of an identical pair to an identical pair (e.g. "CC"->"TT"); (percentage of total adjacent pairs)	42 (6.6%)	13 (92.8%)
Adjacent pairs with non-identical conversions (eg. "AA"->"TC", "AT"->"GC" or "GC" ->"CC") (percentage of total adjacent pairs)	599 (93.4%)	1 (7.2%)

Table S11. Instances of polymorphic adjacent SNPs among the house finch MG strains.

	Position of double SNP in Reference Coordinates													
Strain	14,966	61,514	76,728	120,043	169,641	225,915	241,224	303,492	315,466	572,038	667,905	688,985	715,595	803,438
R Low	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TN_1996		CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
VA_1994	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	AA
KY_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	AA
GA_1995	GG	CC	GG	GG	CC	CC	AA	GG	TT	GG	CC	CC	TC	AA
AL_2001_53	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2001_17	AA	CC	AA	GG	CC	CC	AA	AA	CC	AA	TT	TT	CC	AA
AL_2001_61	GG	CC	GG	AA	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2001_13	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2007_10	GG	TT	GG	GG	TT	TT	TT	GG	CC	GG	CA	CC	TT	AA
AL_2007_05	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
AL_2007_38	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
AL_2007_37	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
TK_2001	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TT	AA
CK_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TK_1998	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TK_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG

Table S12. Counts of unique and total (due to duplication) CRISPR spacers from each strain.

Strain	Total Unique Spacers	Total Spacers
CK_1996	66	75
TK_1998	35	36
TK_1996	93	147
Reference Genome	61	71
TK_2001	38	42
VA_1994	34	36
TN_1995	38	40
KY_1995	35	39
GA_1995	47	50
AL_2001_17	37	37
AL_2001_53	40	40
AL_2001_61	40	40
AL_2001_13	39	39
AL_2007_10	28	28
AL_2007_05	28	28
AL_2007_38	28	29
AL_2007_37	28	28
All Strains	**302**	**805**